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

Martin Beraja
David Y. Yang
Noam Yuchtman*

May 12, 2020

Abstract

Data-intensive technologies such as AI may reshape the modern world. We propose that two features of data interact to shape innovation in data-intensive economies: first, states are key collectors and repositories of data; second, data is a non-rival input in innovation. These characteristics imply an important role for government data in firms' data-intensive innovation. Using comprehensive data on Chinese facial recognition AI firms' software production and government contracts, we identify a causal effect of access to *government* data on *commercial* software production. We then build a directed technical change model to study the state's role in three applications: autocratic states' demand for surveillance AI, data regulation to protect privacy, and industrial policy targeting data-intensive technology. When the degree of non-rivalry is as strong as our empirical evidence suggests, the state can crucially shape the direction of innovation and growth in data-intensive economies.

Keywords: data, innovation, artificial intelligence, industrial policy, regulation, autocracy, directed technical change, political economy
JEL Classification: O30, O40, P00, E22

*Beraja: MIT and NBER. Email: maberaja@mit.edu. Yang: Harvard University and J-PAL. Email: davidyang@fas.harvard.edu. Yuchtman: LSE, NBER, and CESifo. Email: n.yuchtman@lse.ac.uk. Extraordinary research assistance was provided by Haoran Gao, Andrew Kao, Shuhao Lu, Shengqi Ni, Wenwei Peng, Guoli Yin. Many appreciated suggestions, critiques and encouragement were provided by Daron Acemoglu, Ernesto Dal Bó, Ruben Enikolopov, Richard Freeman, Andy Neumeyer, Juan Pablo Nicolini, Maria Petrova, Torsten Persson, Nancy Qian, and John Van Reenen, 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 ("AI" for brevity) technologies hold the potential to transform the modern world. This has attracted a great deal of attention from economists and others.[1] Developing these modern 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 *not* by algorithmic advances, but by access to massive amounts of data.[2]

In this paper, we propose that two features of data interact to shape innovation in data-intensive economies: first, states are key collectors and repositories of data (tellingly, "state" is at the root of the word "statistics").[3] Second, data is a *non-rival* input into innovation: government data accessed to develop software for the state can also be used to develop private sector, commercial products.[4] To examine the empirical relevance of these features of data, we study the prototypical data-intensive sector in which the state has a significant interest: the facial recognition AI industry in China. We document that, following the receipt of contracts to provide services for the state, AI firms produce more software both for *government* and *commercial* purposes when the contracts allow access to greater amounts of government data. This evidence not only establishes the importance of government data in shaping AI innovation, but also provides (to the best of our knowledge) the first empirical estimate of the strength of non-rivalry of data as an input in innovation. We then build a directed technical change model where government data is non-rival across government and private sector innovation. This allows us analyze the role of the state in three applications: autocratic surveillance states demanding AI to monitor their populations, data regulation due to privacy concerns, and industrial policy targeting data-intensive sectors. Disciplining the strength of non-rivalry with our empirical evidence, we show that the state can significantly shape the direction of innovation and growth of data-intensive economies.

[1] For example, recent research has studied the consequences of AI for: business strategy and organization (Agrawal et al., 2018); the existence of jobs and the distribution of income (Acemoglu and Restrepo, 2018; Bessen, 2018; Korinek and Stiglitz, 2017; Sachs, 2018; Stevenson, 2019); industrial organization and market structure (Varian, 2018); as well as the process of R&D and economic growth more broadly (Aghion et al., 2017; Cockburn et al., 2018).

[2] See "Datasets Over Algorithms," by Alexander Wissner-Gross, published by *Edge.org*. Available online at `https://www.edge.org/response-detail/26587`, last accessed November 30, 2019. AI entrepreneur Kai-Fu Lee (former director of Microsoft Research Asia and president of Google China) has even argued that quantities of data are more important than researchers, saying: "the US has 11 times more researchers than China. ... But it is data that is crucial to the implementation of AI technologies, with the higher quantity of data generating better outcomes" (see `https://asiahouse.org/news-and-views/kai-fu-lee-age-ai-china-new-opec/`, last accessed January 14, 2020).

[3] The collection of data is part of the state's fundamental objective of making society "legible" (Scott, 1998).

[4] The non-rivalry of data have been highlighted by Jones and Tonetti (2018), among others.

To conceptualize non-rivalry and guide our empirical strategy, we first present a simple, partial-equilibrium model of data-intensive software production. Taking demand schedules and input prices as given, differentiated firms produce software both for the private sector ("commercial" software) and the government. The *same* government data is an input in the production of both government and commercial software; the marginal productivity of government data in commercial software production corresponds to the strength of the non-rivalry of data (across government and commercial uses). This simple model generates a clear prediction which we can take to the data: an exogenous increase in government data available to a single firm leads to increased production of *both* government and commercial software. Moreover, the ratio of commercial software to government software produced in response to an exogenous increase in government data provides a lower bound on the strength of non-rivalry, which can be estimated even without directly observing the quantity or price of the data to which the firms have access.

A prototypical example of data-intensive innovation in which the state has a significant interest is China's facial recognition AI industry. The development of facial recognition technology requires access to large datasets linking faces to personal identifiers, and the Chinese state collects huge amounts of personally identified data and demands software in order to monitor citizens. Thus, a facial recognition AI firm receiving a government contract to provide surveillance software will gain access to government data (e.g., from surveillance cameras) and be able to accurately match faces to personal records. Therefore, when the degree of non-rivalry of data is high, the firm will be able to to produce new data-intensive software both for the government and for the commercial market. For example, by training or fine-tuning a facial recognition algorithm on a massive government dataset, it can produce a better *commercial* product (e.g., a facial recognition retail payment system).

Given this empirical context, we test the model's predictions and estimate the strength of non-rivalry using a triple differences design. Specifically, we 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. As in an event study design, we compare firms' outcomes — their software releases and patent filings — 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 *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. Specifically, our baseline strategy compares the effects of "data-rich" contracts with public security agencies to those of "data-scarce" government contracts with non-public security agencies. The logic is that public security agencies specifically contract

with AI firms to develop software in order to match surveillance data from cameras to government personal records from entire communities. In contrast, other 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.

To operationalize our baseline strategy, we collect data on the universe of close to 8,000 Chinese facial recognition AI firms' software development and patenting, from *Tianyancha*, a comprehensive database on Chinese firms, licensed by the People's Bank of China, and *PitchBook*, a database on firms and private capital markets around the world. Using a Recurrent Neural Network model with a training set of 13,000 manually labeled items, we categorize software and patents based on whether they are directed towards the commercial market or government use. We merge this data to a separate database of nearly 3,000,000 Chinese government contracts, issued by all levels of the government. We are able to identify contracts awarded for work with a public security agency within the government, which are most likely to provide access to massive, linkable, personal data, collected for monitoring purposes.

We first examine whether receipt of a first data-rich, public security contract increases the number of software products developed by a firm, and whether the increase is *differentially* large relative to the receipt of a data-scarce, non-public security contract. We find that, indeed, in the two years after they are granted, public security contracts generate an additional 1.8 government software products (over and above the effects of a non-public security contract), and an additional 1.5 commercial software products. Findings are similar when examining patents as an alternative outcome. Supporting a causal interpretation of our results, we find *no* differential software production levels or pre-treatment trends associated with receipt of an initial government contract of either type. Our estimated effects of a government contract on both private and public sector innovation allow us to bound the non-rivalry parameter in our partial equilibrium model by estimating the ratio of the elasticities of private and public innovation with respect to receiving a contract. Importantly, the ratio of the commercial to government software production response indicates a degree of non-rivalry of data significantly greater than the full rivalry benchmark of 0.

Note that a key empirical challenge when studying the economic role of data is that it is difficult for researchers to systematically observe the data possessed by firms. Our comparison of public security contracts to other government contracts exploits one margin along which firms differ in their access to data. As an alternative approach, we also examine variation *within* the set of public security contracts: some public security agencies

possess much larger monitoring apparatuses than others. As a proxy for an agency's *ex ante* monitoring capacity, we collect data on the number of surveillance cameras in each Chinese province. We then test whether public security contracts differentially increase new product and patent development in localities with above-median monitoring capacity. The logic is that a contract with a public security agency in a province with more surveillance cameras will be particularly "data-rich", plausibly providing firms with access to larger datasets. Consistent with data playing a causal role in the production of new software, we find that firms receiving contracts with public security agencies that possess more monitoring capacity produce differentially more commercial and government software following receipt of the contract. Again, we find no evidence of differential pre-contract software production levels or trends, supporting a causal interpretation of our results.

Next, we provide a range of corroborating evidence for our proposed mechanism of access to government data contributing to product innovation. One would expect that 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). Importantly, these data-complementary software products are *distinct* from the *AI* software studied above. Indeed, we find that data-complementary software production significantly, and differentially, increases after firms receive public security contracts. Yet, this is only so on average: some firms receiving public security contracts do not produce any data-complementary software at all. Thus, we next use the production of data-complementary software as indicating that certain public security contract were particularly data-rich. We find that, within firms receiving public security contracts, those who produce data-complementary software following the receipt of the contract — compared to those who do not — differentially produce more government and commercial AI software. Finally, we test whether the most data-intensive output — in the case of facial recognition AI, this is facial recognition technology that matches identity in *videos* — increases following receipt of a data-rich contract. We find that this is the case. Moreover, we find that firms that produce video software for the government — those firms that plausibly were given access to the largest quantities of government data — exhibit differentially larger increases in commercial video AI software as well.

We are also able to rule out several plausible alternative mechanisms. First, one might wonder whether firms with particular characteristics (e.g., pre-contract software production or pre-contract financial circumstances) would have exhibited increased software production even in the absence of receiving a data-rich government contract. To address this concern, we control for a range of firm characteristics interacted with time period fixed effects, thus allowing firm differences to flexibly affect software production over time.

Including these controls does not affect our results. A second concern is that data-rich contracts differ from data-scarce contracts on dimensions other than data access, which might also affect software production. To address this possibility we add to our baseline empirical specification controls for contract characteristics — e.g., the monetary value of the contract, and the GDP of the province where the contract was issued — interacted with time period fixed effects. These controls, too, do not affect our results. A related concern is that contracts provide firms with credible signals of quality, and these signals may differ systematically by type of contract. We thus examine the effects of *second* contracts which, while estimated on a selected population, arguably allow us to estimate the effect of a data-rich contract in a context where signaling should be less important. We find very similar effects of second contracts on software production, suggesting that signaling likely does not drive our results. Finally, we account for the possibility of learning by doing under a government contract playing an important role by directly controlling for firms' lagged software production; this has very little effect on our estimates. A broad range of evidence thus suggests an important role of access to non-rival government data in shaping the development of AI innovation.

We conclude the paper by analyzing how the two key features of data — states being among the key collectors and repositories of data, and data being a non-rival input in the production of data-intensive technologies — interact to shape growth and the direction of innovation in data-intensive economies.

To do so, we embed our partial equilibrium model of data-intensive software production into a general equilibrium directed technical change model along the lines of Acemoglu (2002). Innovator firms can develop and supply differentiated varieties of data-intensive government and private (commercial) software, as well as other, non-software, varieties which do not use data. Private software and non-software are intermediates inputs in the production of a final good consumed by a representative household. Government software varieties are instead used to produce "surveillance and monitoring services" which are purchased by the state. The crucial difference in software production technologies is that non-rival government data is the only data input for producing government software, whereas it is a gross substitute with private data in commercial software production. Moreover, we assume that while private data is excludable and can be purchased in the market, government data can only be accessed by signing contracts with the state and producing government software for it — analogous to our empirical setting. Therefore, our economy features three types of firms: those innovating and producing government software and thus being able to produce private software using both the non-rival government and private data; those that do not produce government software and thus produce private software using private data alone; and, those that produce non-data-intensive

products for the alternative sector. Moreover, we assume that government (private) data is produced as a by-product of surveillance and monitoring services (private transactions as measured by output of the final good). Finally, we let the state choose a policy involving: a level of surveillance expenditures, government data provided to firms, taxes on private data production, and lump sum transfers to the household. Given a state policy, we then characterize conditions under which a balanced growth path equilibrium with free entry of innovator firms exists and is unique.

Next, in three applications, we consider alternative state policies that capture different dimensions along which states may vary: first, the development of autocratic surveillance states that monitor and control citizens; second, laws and norms regarding the protection of privacy and the repugnance of markets for data; and third, the state's engagement in industrial policy to promote data-intensive innovation. While admittedly speculative, the model allows us to illustrate the similar economic mechanism driven by data's non-rivalry that underlies the role of the state across these applications. In particular, in our baseline model, we show that *only* policies affecting the supply of *government* data to firms affects the growth rate and direction of technical change. This means that surveillance states can grow more quickly than states respecting civil liberties; it means that privacy restrictions only matter as they relate to government data; and, it means that traditional industrial policy (e.g., subsidizing private data purchases) will be ineffective.[5]

In addition, the model gives a structural interpretation to our reduced-form evidence on non-rivalry and therefore provides a sense of the magnitudes involved in the impact of the state in each of the applications. Our parameterized model implies that, compared to the US, an autocratic surveillance state like China, which collects and shares more government data, could exhibit a greater growth rate by between 0.6 and 0.95 percentage points, and bias its innovation towards private AI software by between 7.8 and 20.9 percent. This "autocracy advantage" in growth is large enough to offset an "autocracy tax" on private firms of up to 11.5%. As an alternative, consider a state that strongly protects privacy, like Germany: our model estimates that decreasing the number of surveillance cameras from the US level (15.28 per 100 residents) to Germany's (6.27 per 100 residents) would decrease the growth rate by between 0.4 and 0.9 percentage points and bias innovation away from private AI software by between 5 and 22.7 percent. Finally, consider the effects of industrial policy: as noted above, traditional industrial policy does not affect the equilibrium growth rate on the balanced growth path in our baseline model. Yet, even in our extended model where subsidies to private data are non-neutral, we find that (under most parameterizations) the same increase in the economy's growth rate can be achieved at a lower

[5]We relax the baseline model to allow for "non-neutrality" of private data, and find that policies affecting the supply of government data will still generally have much larger effects than policies affecting private data alone.

fiscal cost through industrial policy on the provision of government data. Our model thus suggests that state policy could play a critical role in shaping the global landscape in the coming decades as data-intensive technologies become widespread.

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. We identify a crucial characteristic of data that shapes the impact of the state on innovation: data's non-rivalry across uses. Our analysis complements a recent literature studying the effects of specific characteristics of information on innovation: Williams (2013) studies the non-excludability of government research on genes (in contrast with the excludability of private sector research); Aghion et al. (2017) and Jones and Tonetti (2018) study non-rivalry of data across firms. We instead study the impact of non-rivalry across public and private applications. In so doing, we provide the first empirical estimate of the strength of non-rivalry of data across different uses.

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 innovation policy also suggests an important role for the state in encouraging R&D — see Bloom et al. (2019) for a review.[7] We make three primary contributions to these literatures. First, we study a new context: the effects of the state on the development of modern AI innovation, a technology at the current frontier, 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 the non-rivalry of data across public and private applications could generate consequences similar to those

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

[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.

achieved by industrial policy and innovation policy, despite the incidental nature of the state's engagement, as in Slavtchev and Wiederhold (2016).[8] Third, our macroeconomic analysis reveals important, unexpected implications for industrial policy and innovation policy in data-intensive economies. We find that when data are sufficiently non-rival (as suggested by our empirical analysis), traditional industrial policy favoring data-intensive technologies — specifically, subsidies to purchase private data — will be less effective than policies that provide firms with access to government data.

By placing our analysis of AI innovation within a model of directed technical change, our work contributes 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) 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). 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 procuring 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 particularly valued — and thus accumulated — by modern autocratic states. In addition, a fundamental aim of AI technology — to make accurate predictions — is aligned with the surveillance and social control agenda of autocratic states. 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. The risk of *ex post* taxation or expropriation of entrepreneurs will mean *ex ante* less investment (North, 1991; Acemoglu and Robinson, 2012).

[8]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).

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). In the domain of AI technology, autocratic states' political incentives, along with the non-rival nature of data as an input, facilitate data-intensive innovation even for private applications — the shadow cost of data is substantially lower and the alignment between the state and private sector could resolve the expropriation risks and commitment problems traditionally faced by private entrepreneurs under autocracy.[9] Our analysis thus helps resolve 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.[10]

In what follows, we will present the partial equilibrium framework on data-intensive innovation in Section 2, 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 3; then the empirical strategy and results in Section 4. In Section 5, we embed the partial equilibrium model of data-intensive software production into a general equilibrium framework of directed technical change. We discuss three applications of the framework in Section 6, and offer concluding thoughts in Section 7.

2 Partial equilibrium framework on data-intensive innovation

This section presents a model of a firm producing data-intensive software for both the government and private sector. We use it to conceptualize government data non-rivalry and derive observable predictions and parameter estimation strategy to guide our empirical exercise in the next section.

Let the technologies for producing software for the private sector (q_p) and software for

[9] 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 data's non-rivalry) in order to preserve its *economic* rents.

[10] A large literature studies the Chinese economy and its spectacular growth in 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 (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.

the government (q_g) be:

$$q_g = \left(d_g\right)^{\beta} \left(X(\overrightarrow{x_g})\right)^{1-\beta}$$
$$q_p = \left(h(d_g, d_p)\right)^{\beta} \left(X(\overrightarrow{x_p})\right)^{1-\beta}$$

where d_g is government data, d_p is data from other sources, and $\overrightarrow{x}$ is a vector of other inputs.

Definition 1 (Non-rivalry) *The degree of non-rivalry of government data is:*

$$\sigma \equiv \frac{\partial log(h)}{\partial log(d_g)}$$

The definition implies that the degree of non-rivalry of government data is higher whenever its marginal product in private software production is higher. In other words, the non-rivalry is defined at the level of within-firm-across-uses.

Next, we consider a firm facing demand schedules for both its private and government software with a constant elasticity $\chi > 1$. Given government data d_g and a vector of input prices $\overrightarrow{W}$, it chooses d_p, $\overrightarrow{x_p}$, $\overrightarrow{x_g}$ to maximize profits.

Proposition 1 (Implications of non-rivalry) *Let $\Delta log(d_g) > 0$ be an exogenous increase in firm access to government data. Let $\Delta log(q_p)$ and $\Delta log(q_g)$ be the resulting change in private and government software production. Assume that $h(d_g, d_p)$ is a CES function with elasticity of substitution γ and $X(.)$ is constant returns to scale. Then,*

$$\Delta log(q_g) = \frac{\chi\beta}{1 + (\chi - 1)\beta} \Delta log(d_g)$$
$$\Delta log(q_p) = \frac{\chi\beta\sigma}{1 + (\chi - 1)\beta + \gamma(1 - \sigma)} \Delta log(d_g),$$

and

$$\sigma = \frac{\frac{\Delta log(q_p)}{\Delta log(q_g)}}{1 - \frac{\gamma}{1 + \beta(\chi - 1) + \gamma} \left(1 - \frac{\Delta log(q_p)}{\Delta log(q_g)}\right)}$$

Proof. See Appendix A.1. ■

Proposition 1 provides two empirically testable predictions. First and most simply, the increase in access to government data causes government software production to increase. Second, the increase in access to government data leads to the increase of the commercial software production, if and only if government data is non-rival ($\sigma > 0$) and at least imperfectly substitutable with private data ($\gamma < \infty$).

Proposition 1 also guides an empirical strategy to estimate the non-rivalry parameter σ, even in the absence of precise information on the quantity and price of data to which the firms have access. Qualitatively, everything else equal, the degree of non-rivalry is higher whenever one observes a higher elasticity of private to government software changes ($\frac{\Delta log(q_p)}{\Delta log(q_g)}$). Quantitatively, based on the last expression, this elasticity is a lower bound for the degree of non-rivalry.

3 The government and China's facial recognition AI: context and data sources

3.1 Empirical context

China's facial recognition AI sector is a prototypical setting in which to examine the impact of access to government data on innovation and to estimate the strength of non-rivalry of data as an input in the innovation process. First, because facial recognition AI is extremely data-intensive: the development of the technology requires access to large datasets linking faces to personal identifiers. Second, because the Chinese state, to a greater extent than most others, collects huge amounts of personally identified data and demands software in order to monitor citizens.

This context allows us to empirically evaluate the theoretical predictions of our model. Private firms receiving a procurement contract to provide facial recognition services to the Chinese government implicitly receive access to large quantities of government data. To the extent that data is a non-rival input, the government data could be used for *commercial* innovation and product development, beyond the production of new software for the government. We exploit within-firm variation over time in the receipt of procurement contracts to estimate the effect of access to government data on government and commercial software production, as well as to estimate the degree of non-rivalry of data in the production function.

3.2 Data sources

To conduct our empirical analysis, we construct a novel dataset combining three elements: first, a (nearly) complete list of Chinese firms producing facial recognition AI; second, comprehensive information on government contracts; and, third, technological innovations, as measured by new products and patents. We describe these elements of our data in turn.

Chinese facial recognition AI firms We identify (close to) all active firms based in China producing facial recognition AI using information from *Tianyancha*, a comprehensive data-

base licensed by the People's Bank of China (i.e., China's central bank) on Chinese firms (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.[11]

In addition to the firms' roster, 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.

Chinese government contracts 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). Our empirical analyses focus on government contracts involving public security, as they potentially provide access to large datasets collected by the government for monitoring purposes. We capture the following four types of public security contracts: *(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 administrative units for domestic security and stability maintenance, the government's political and legal affairs commission, and various "smart city" and digital urban management units of the government. We identify 410,510 such public security contracts. We will also examine procurement contracts unrelated to public security (e.g., contracts with schools to monitor cheating during exams), which can be seen as "placebo treatments" that indicate firms' with relationships with the government, but without providing access to large amounts of personally identified facial data.[12]

[11] Note that this combines subsidiaries owned by the same mother firm. 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).

[12] We present the cumulative number of procurement contracts of each type in Appendix Figure A.4, 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 are steady increasing since 2013. There is strong seasonality in the month during

In order to identify public security contracts procuring facial recognition AI, we match the public security contracts with the list of facial recognition AI firms, identifying 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 also identify "placebo" non-public security contracts procuring facial recognition AI. 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.

Table 1 presents summary statistics for various procurement contracts: those that involve public security arms of the government (Columns 1 and 2), and those that are related to non-public security (Columns 3 and 4). Overall, public security contracts have bigger monetary size that non-public security ones, though the latter tend to be signed with higher administrative units of the government (e.g., provincial government or above), and they tend to emerge around the same time. We will explicitly examine the role of monetary size of the contracts in Section 4.2.

Facial recognition AI software products and patents We collect all software registration records for our facial recognition AI firms from China's Ministry of Industry and Information Technology, with whom Chinese firms are required to register new software releases and major upgrades. Patent filings for our facial recognition AI firms are collected from the *Tianyancha* database. We are able to validate our measure of software releases and patent filings for a single large firm, cross-checking our data against the IPO Prospectus of MegVii, the world's first facial recognition AI company to file for an IPO. We find that our records' coverage is comprehensive (at least in the case of MegVii): MegVii's IPO Prospectus lists 302 important patents, and our records contain 1,187 patents owned by MegVii, of which the 302 disclosed important patents are a subset; the Prospectus contains 103 software releases, all of which are included in our dataset. For the primary empirical analyses, we focus on software releases, but all of our results are robust to examining patent filings instead (results reported in the Appendix).

Categorizing software products and patents We use a Recurrent Neural Network (RNN) model with tensorflow — a frontier method for analyzing text using machine learning — to categorize the software products and patents according to their intended customers and (orthogonally) by their function. We are thus able to distinguish between software products and patents developed for the government (e.g., "smart city — real time monitoring

which the contracts are signed: we s peaks in December and a dip in the subsequent January. This is likely due to the administrative budget cycles.

system on main traffic routes") and software products and patents developed for commercial applications (e.g., "visual recognition system for smart retail"). We allow for a residual category of general application software or patents whose description does not clearly specify the intended user (e.g., "a synchronization method for multi-view cameras based on FPGA chips").[13] We also distinguish between software and patents 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 AI-complementary, involving data storage, data transmission, or data management (e.g., "a computer cluster for webcam monitoring data storage"). Finally, we separately identify within the category of AI-related software and patents 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.[14]

To implement the categorization using the RNN model, we manually label 10,000 software products and patents to produce a training corpus. We then use word-embedding to convert sentences in the software and patent 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 86.8-88.3% accuracy in categorizing software and patents in the validation data. Appendix Tables A.2 to A.5 present the top words (in terms of frequency) used for the categorization; Appendix Figures A.6 and A.7 present the density plots of the algorithm's category predictions for software and patents respectively; Appendix Figures A.8 and A.9 show the summary statics of the categorization output by customers and by functions for software and patents respectively; and finally, Appendix Figures A.10 and A.11 presents the confusion matrix (Type-I and Type-II errors) of the prediction relative to categorization performed by humans for software and patents respectively.[15][16]

Table 2 presents summary statistics for all facial recognition AI firms (Column 1), those

[13]By coding as "commercial" only those products and patents that are explicitly and specifically linked to commercial applications, and excluding products and patents with ambiguous use, we will estimate a more conservative lower bound on the degree of non-rivalry.

[14]The differences in 1-to-1, N-to-1, and N-to-N matching in facial recognition are some times referred to as facial authentication versus facial recognition; see Gates (2011) for more details.

[15]The algorithm is very accurate in categorizing software and patents for government purposes. The algorithm is relatively conservative in categorizing software products and patents as for commercial customers, and relatively aggressive in categorizing them as for general purpose. This downwardly biases our measures of the relative increase in commercial innovation and production relative to government ones.

[16]See Appendix Figure A.12 for prediction accuracy by years when software and patents were released/registered. There is no evidence that the prediction accuracy rates differ across years.

with at least one public security procurement contract (Column 2), and those without any public security procurement contracts (Column 5). One can see that, on average, firms with at least one public security procurement contract are much larger in capitalization, and have produced more software products and patents. Accounting for differences (both observable and unobservable) between the firms receiving public security contracts and those that do not will be crucial to identify the effect of a government contract.

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

4.1 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 and patent filings — 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 *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. Some contracts — particularly those awarded by public security agencies, and those awarded in provinces with greater monitoring capacity — are likely to be "data rich," while other contracts will be "data scarce."

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

$$y_{it} = \sum T_{it} \times Data_i + \sum T_{it} + \alpha_t + \gamma_i + \epsilon_{it}.$$

The outcome variable, y_{it}, is the cumulative number of software releases by firm i up to the semi-year period t (as an alternative outcome, we consider firms' patent filings analogously to our study of software; results presented in the Online Appendix). 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 government contract, and $Data_i$, a dummy variable indicating whether the firm's first government contract was data-rich.[17]

[17] We focus on the effect of the initial contract, because the receipt of subsequent contracts is endogenous to

In practice, we operationalize $Data_i$ in two ways: first, we split government contracts depending on whether the award was from a government agency providing public security (data-rich) or not (data-scarce). Second, as an alternative (presented in the Online Appendix), we split public security contracts into those awarded in provinces with above-median video camera monitoring capacity (data-rich) or below-median (data-scarce).[18]

Take as an example of 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 that the firm obtains access to. In contrast, a non-public security contract 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 involved AI firm obtains access to relatively small amount of data compared to the public security ones.

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 firms, 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.

Threats to identifying the causal effect of receiving access to government data must vary both across time and across firms in a manner correlated with the type of contract received and with firms' software or patent production. For example, firms with a particular set of *ex ante* characteristics may be more likely to receive a data-rich government contract at a particular time, and these characteristics may be associated with more software production in subsequent periods regardless of whether the firms actually received the contracts. Another concern is that our proxies for data-rich contracts are correlated with some other contract characteristic that might drive subsequent software or patent production.

firms' performance in their initial contracts (Appendix Figure A.5 plots the arrival of procurement contracts relative to the time the first contract is received).

[18] It is important to note that surveillance capacity is not perfectly correlated with local market size or wealth. For instance, while the Shanghai Municipal Government represents the richest locality in China in terms of per capita GDP, the existing surveillance capacity in Shanghai (per capita CCTV cameras), is dwarfed by regions that presumably present greater public security risk (e.g., Xinjiang, where the Uighur minorities reside, and Guangdong, where large numbers of migrant workers reside). Provinces with above-median monitoring capacity are Xinjiang, Beijing, Guangdong, Anhui, Shandong, Fujian, Zhejiang, Shaanxi, and Sichuan. Appendix Table A.1 presents summary statistics of localities with above and below median surveillance capacity.

While we cannot entirely rule out the possibility of time×firm-varying shocks with non-experimental data, we *can* conduct a set of falsification exercises by testing whether $\sum T_{it} \times Data_i$ differ from zero *prior* to contract receipt; that is, conducting a test of parallel pre-treatment trends. We can also control for a range of firm characteristics interacted with time period fixed effects to test whether *ex ante* firm characteristics, rather than the contract, seem likely to drive differential software or patent production over time. Finally, we can control for contract characteristics other than data access (e.g., the contract's monetary value and characteristics of the contract granting agency) interacted with time period fixed effects to evaluate the role of non-data contract characteristics in driving our results. We provide evidence for our proposed mechanism and evaluate alternative mechanisms in detail, below.

4.2 Results

We first present the baseline results in Section 4.2.1 following the empirical strategy described above; we the describe a variety of supporting evidence that highlight the role of government data in firms' innovation in Section 4.2.2; finally, in Section 4.2.3 we discuss alternative hypotheses and present evidence to rule them out.

4.2.1 Baseline estimates and robustness checks

We first estimate our baseline specification, comparing the effects of public security contracts (data-rich) and non-public security contracts (data-scarce) on firms' production of software. In Figure 1, top panel (blue circles), one can see that following receipt of a public security contract, there is a statistically significant increase in both government and commercial software. Importantly, prior to the receipt of the contract, there is no differential pre-contract level or trend rate of software production. One can also see in Figure 1 (black squares) evidence suggesting an important role of access to government *data*: following receipt of a data-scarce non-public security contract, increases in software production are much smaller than following receipt of a data-rich public security contract. One can see the coefficient estimates and standard errors from the baseline regressions in Table 3, Columns 1 and 2.

Our baseline estimates indicate that receiving a data-rich government contract (relative to a data-scarce one) increases a firm's number of government software products by around 1.8 over 3 semi-years, and increases a firm's number of commercial software products by also 1.5 over 3 semi-years. These estimates allow us to bound the non-rivalry parameter in our partial equilibrium model. Specifically, we estimate the ratio of the elasticities of commercial software releases and government software releases with respect to

receiving a public security procurement contract (differencing out the elasticities with respect to receiving a non-public security contract to isolate the dimension of data access). The ratio of the commercial software response to the government software response indicates a non-rivalry of data parameter bounded below by 0.83 (= 1.5/1.8), significantly greater than the full rivalry benchmark of 0.[19]

The patterns observed in Figure 1 and Table 3 are confirmed considering a different measure of firms' innovative output — patents — or an alternative classification of data-rich and data-scarce contracts — comparing public security contracts in provinces with high or low monitoring capacity. In Appendix Figures A.13, A.14, and A.15, and Appendix Table A.6, one can see that receipt of a data-rich contract is consistently followed by a significant increase in the production of government-oriented products and innovation; it is also followed by an increase in the production of *commercially*-oriented products and innovation.[20] In none of the analyses does one observe differential pre-contract software production or patenting, supporting a causal interpretation of our findings.

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 bureau could be meant to provide security service just for the government office building). This has no impact on our results (see Appendix Table A.8). 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 and patents — timestep, embedding, and nodes. We vary these three parameters, re-configure the RNN LSTM algorithm, re-categorize software and patents, and re-estimate the baseline empirical specification. We show that these algorithm parameter choices have no impact on our results (see Appendix Table A.9). Moreover, 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 the narrow window (studying a balanced panel over too long a window substantially reduces the sample size). These changes, too, have no impact on our findings (see Appendix Table A.10). Finally, we control for mother-firm's fixed effects, or we aggregate firms in our sample to the mother-firm level, studying the impact of the first contract received by the mother firm on software produced. These analyses again produces results that are very similar to our

[19]To the extent that certain software products could be intended for commercial applications but are categorized as general purpose software because the commercial intents are not explicitly stated in the software description, we estimate a conservative lower bound of the non-rivalry parameter.

[20]In addition, we combine the three tiers of data-rich and scarce contracts — non-police contracts, police contracts with locality with low surveillance capacity, and police contracts with locality with high surveillance capacity — in one regression specification. The results are presented in Appendix Table A.7.

baseline estimates (see Appendix Table A.11).

4.2.2 Evidence on the importance of data as a mechanism

The AI firms in our sample produce both more government software and more commercial software just after receiving contracts with government agencies that provide them with access to greater amounts of data. They were not differentially producing software just before receiving these contracts. This suggests an important role of access to data in firms' production of new products and innovation. In this section we provide further supporting evidence for our proposed mechanism.

First, we examine whether firms receiving data-rich government contracts differentially produce data-*complementary* software. One would expect that 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). Importantly, these data-complementary software products are *distinct* from the *AI* software studied above. In Figure 1, we present our baseline estimates, but focusing on the outcome of software products that we categorize as data-complementary. Consistent with government contracts providing quantities of data that go beyond firms' existing data management capacities, we find that the production of such data-complementary software significantly increased following receipt of a data-rich contract (relative to receipt of a data-scarce contract).

Because data-complementary software can be seen as an input into subsequent product development (i.e., complementing the newly acquired data), it is natural to test whether firms that both receive data-rich contracts *and* develop data-complementary software disproportionately increase their AI software production.[21] Indeed, we find this to be true: in Figure 2, one can see that AI software production — both for government and commercial uses — increases dramatically, specifically among firms that develop data-complementary software and patents after receiving their first government contacts.

As an additional test of our proposed mechanism, we examine whether, within the set of AI software, one observes significant effects on the most data-intensive output: in the case of facial recognition AI, this is facial recognition technology that matches identity in *videos* (as opposed to merely photos). Suggesting the importance of access to government data, we find that video-related AI software *is* increasingly produced after firms receive a data-rich government contract (see Figure 1, bottom right panel, and Table 3, Column 4). In fact, video-related software represents a large share — over 50% — of the new technologies developed post-contract.

[21] The production of data-complementary software and patents might also be an empirical proxy for firms' receiving access to particularly large quantities of data — this would be another reason why firms exhibiting such production should differentially increase their AI software production.

Finally, we consider whether firms that produce video software for the government — those firms that plausibly were given access to the largest quantities of data — exhibit differentially large increases in commercial video AI software. Indeed, in Appendix Table A.12, Panel A, we present evidence of particularly large effects among this subset of firms. One might wonder whether this could reflect learning by doing, rather than access to data; however, we find large effects *outside* commercial video AI production (Appendix Table A.12, Panel B). We also see effects that are equally large regardless of whether a firm had produced video AI software prior to receiving its first government contract, again suggesting that access to data, rather than learning how to produce AI software, drives these results (we consider learning by doing more generally below).

4.2.3 Evaluating alternative hypotheses

While a range of analyses suggest an important role of access to government data in shaping firms' production of AI software, it is important to consider alternative mechanisms.

Potentials and catch-ups First, one might wonder whether firms with particular characteristics select into data-rich contracts; perhaps these firms would have exhibited different software production dynamics even in the absence of receiving a contract. For example, some firms with little pre-contract software production — but having significant potential — may receive their government contracts, then exhibit "catch up", significantly increasing software production post-contract. To address this possibility, we add to our baseline specification interactions between a firm's pre-contract software production and the full set of time period fixed effects (formally $\sum T_{it} \times software_i$). Adding this set of controls has no impact on our results (see Appendix Table A.13, Panel A).[22]

Provision of capital Another set of alternative hypotheses involve contracts affecting 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, 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 $\sum T_{it} \times value_i$). We add these interactions to our baseline specification and find that they do not affect our results (see Appendix Table A.13, Panel D). Second, we can examine whether data-rich contracts differentially benefit firms that were relatively poorly-capitalized prior to receipt of the first contract. We split the sample into firms with above or below median amounts of external financing prior to receiving their first contract, and estimate our baseline specification

[22] This analysis also helps address concerns about government contracts generating economies of scale.

on these two subsamples. We find that, if anything, the effects of data-rich contracts are greater for the *well*-capitalized firms, again suggesting that contracts' provision of capital does not drive our results (see Appendix Tables A.15, Panel B, and A.16, Panels B and C).

Access to bigger markets Another possibility is that our estimated effects of data-rich contracts in fact capture the effects of producing in rich commercial markets.[23] 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 $\sum T_{it} \times market_i$). We add these interactions to our baseline specification and find that they do not affect our results (see see Appendix Table A.13, Panel E).

Indeed, our results are robust to controlling for a full set of time-period fixed effects interacted with a vector of firm characteristics (formally $\sum T_{it} \times firm_i$), as well as with a vector of contract characteristics other than our proxies for data content (formally $\sum T_{it} \times contract_i$), see Figure 1, Table 3, Columns 5–8, and Appendix Table A.13, Panel F).

Government contracts as signals One might wonder whether the signaling value of the contract plays a role in our results: perhaps firms receiving a government contract attract additional external funding or human capital, which contribute to the production of software and patents. Again, our differencing out of data-scarce contracts' effects should alleviate this concern. As an alternative approach, we can examine the effects of *second* contracts: 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 or patent production around the time of receiving a second contract, we continue to find significant effects of receiving a data-rich contract on software and patent production (see Appendix Table A.14).[24]

Learning by doing Finally, it is possible that data-rich contracts generate more AI software and patenting not because of the data they provide, but because of firms' opportunities for learning by doing under the contract. To address this concern, we estimate our baseline specification, but now including lagged (not only pre-contract) values of firms' AI software production (either government software, commercial software, or total). If learning by doing plays an important mediating role in driving our findings then controlling for lagged software production should significantly drive down the estimated impact of receiving a data-rich contract. We find that controlling for lagged software production does not significantly affect our findings (see Appendix Table A.15, Panels C and D).

[23] Costinot et al. (2019) show that innovation is responsive to characteristics of the local marketplace.

[24] To make this test distinct from our analysis of first contracts, we limit our sample to second contracts received at least 6 months after the first.

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 and patents, but also commercial software products and patents. This is possible due to the non-rivalry of government data in firms' production functions. We next explore the macroeconomic implications of these findings.

5 A model of data-intensive technical change

Having shown that access to government data plays an important role in the firms' production of data-intensive innovation, we now embed our partial equilibrium firm model of data-intensive software production (described in Section 2) into a general equilibrium directed-technical model along the lines of Acemoglu (2002). Our goal is to analyze how growth and the direction of innovation in data-intensive economies is shaped by two features of data: states being key collectors and repositories of data, and data being a non-rival input in innovation. This section presents the environment, defines a "state policy," and characterizes a balanced-growth path equilibrium for a given state policy. The next section analyzes the implications of alternative state policies in three applications.

Model overview We model an economy in which firms innovate to develop and supply differentiated varieties of government and private (commercial) software — which require data in production — as well as other, non-software, varieties — which do not. Private 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.[25] Government software varieties are intermediate inputs into the production of a government good. The state purchases this government good, which we refer to as "surveillance and monitoring."

As in our partial equilibrium framework, we assume that government data is non-rival across uses: that is, government data is a gross substitute with private data for producing private software. Crucially, government data is necessary for producing government software (we assume for simplicity that private data cannot be used to produce government software).

Government data is supplied by the state and is produced as a by-product of surveillance and monitoring. Private data is supplied by a representative firm as a by-product

[25] In the next section, we also consider a household that has preferences over surveillance and monitoring as well as over the amount of data in the economy due to privacy concerns.

of all private transactions in the economy as measured by total output of the final good.[26] Furthermore, while both types of data are excludable, we assume that only private data can be purchased in the market. Instead, government data can only be accessed by obtaining a contract for producing government software varieties for the state.

Therefore, the state chooses a policy that involves: a level of expenditures on surveillance and monitoring (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 transfers to households and distortionary taxes (potentially on private data production). Note that these choices are not independent of each other. Given state policy, potential entrants can choose to innovate on and supply new varieties of government software, private software, both types of software, or only non-software varieties. Firms will innovate and enter such that, in an equilibrium on the balanced growth path, 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: private (commercial) software Y_p, 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 produce a final good Y using a CES aggregator:

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

We again assume the firm is competitive and maximizes static profits given sectoral prices

[26] 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 firms, 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.

p_p and p_z, and the price of Y which we normalize to 1. This implies that prices satisfy:

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

Innovators A software variety j is supplied by a monopolist "innovator." As in Section 2, we assume that producing software of a higher quality is data-intensive.[27] Dropping the j index for notational convenience, government software production uses government data d_g and innovator effort x_g to produce a variety of quality q_g. Private software production uses both government and private data d_g, d_p as well as effort x_p to produce a variety of quality q_p.

Assuming the production technology $h(d_g, d_p)$ introduced in Section 2, the firms produce government and private software using the following technologies:

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

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

where $\alpha < 1$ governs the relative productivity of government vis-à-vis private data, and $\gamma > 1$ describes their gross substitutability in private software production.[28]

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 the marginal cost of effort is ϕ units of the final consumption good and that all product varieties fully depreciate after use.[29]

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

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

$$\Pi_p(d_g, p_p, p_d) = \max_{x_p, d_p} p_p q_p(d_g, d_p, x_p)^{1-\frac{1}{\chi}} - \phi x_p - p_d d_p, \tag{8}$$

with a corresponding private data demand schedule $d_p(d_g, p_p, p_d)$.

We then consider how new varieties are introduced. We assume that innovators can

[27] 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.

[28] The assumption of gross substitutability is important because, as we will see later on, it allows innovators to produce private software even without access to government data.

[29] 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.

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 non-rivalry of government data and the fact 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 private 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 government data $\bar{d}_g$. The innovator can then choose to also introduce a new private software variety using the non-rival government data in its production. Finally, if the innovator does not obtain the government contract, it can again choose to introduce a new private 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 private software variety innovators. Given the assumptions above and setting $\mu_g = \mu_p = 1$, a balanced growth path equilibrium where both types of software producing firms are present requires:

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

$$1 = \frac{\Pi_p(0, p_p, p_d)}{r} \tag{11}$$

Finally, for non-software innovators, the R&D investment yields new varieties with quality $q_z = x_z^{1-\beta}$, where x_z is again effort. 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 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:

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

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 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 *ad valorem* taxes τ on aggregate private data spending $p_d D_p$ and lump sum taxes T on households such that budget balance holds.

$$p_g G = \tau p_d D_p + T. \tag{16}$$

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. Then, given a measure N_g of government software innovators and a dataset available to them $\bar{d}_g$, we have that:

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

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 2 (State policy) *A state policy is a dataset available to government software innovators $\bar{d}_g$, government software expenditures for surveillance purposes relative to private output $p_g G/Y$, private data taxes τ, and lump sum taxes relative to private 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. The final goods representative firm produces D_p as a by-product of private transactions as measured by aggregate output Y. With data mining effort u it can extract $(\kappa_p Y)^{\psi} u^{1-\psi}$ units of data out of Y. Given unit cost of effort and price of data net of taxes $p_d(1-\tau)$, this implies that the supply of private data is:[30]

$$D_p = \left((1-\psi)(p_d(1-\tau))\right)^{\frac{1-\psi}{\psi}} \kappa_p Y. \tag{18}$$

Equilibrium We now consider a balanced growth path equilibrium (BGP) where all variables grow at constant rate η. From the representative household's Euler equation, we obtain that in a BGP:

$$r = \theta\eta + \rho \tag{19}$$

Moreover, we require the following market clearing conditions in the goods and data markets:

$$\tilde{N}_p q_p(0, p_p, p_d)^{1-\frac{1}{\chi}} + N_g q_p(\bar{d}_g, p_p, p_d)^{1-\frac{1}{\chi}} = Y_p = \left(\frac{p_p}{1-a}\right)^{-\epsilon} Y \tag{20}$$

$$N_z q_z(p_z)^{1-\frac{1}{\chi}} = Y_z = \left(\frac{p_z}{a}\right)^{-\epsilon} Y \tag{21}$$

$$N_g q_g(\bar{d}_g, p_g)^{1-\frac{1}{\chi}} = Y_g = G \tag{22}$$

$$\tilde{N}_p d_p(0, p_p, p_d) + N_g d_p(\bar{d}_g, p_p, p_d) = D_p = \left((1-\psi)(p_d(1-\tau))\right)^{\frac{1-\psi}{\psi}} \kappa_p Y \tag{23}$$

$$N_g \bar{d}_g = D_g = \kappa_g G \tag{24}$$

where $\tilde{N}_p$ is the total number of private software varieties produced by firms without a government contract and N_g is the number produced by firms with a government contract (which is also the number of government software varieties), and N_z is the number of non-software varieties.[31]

Definition 3 (BGP Equilibrium) *Given a state policy $\{\bar{d}_g, p_g G/Y, \tau, T/Y\}$, a balanced-growth path equilibrium is a set of prices $\{p_p, p_z, p_g, p_d, r\}$, relative varieties $\tilde{N}_p/N_z$ and N_g/N_z, and*

[30] This also implies that the marginal revenue from an additional unit of output is $P + \kappa_p \psi (1-\psi)^{\frac{1-\psi}{\psi}} (p_d(1-\tau))^{\frac{1}{\psi}}$, where P is the price at which goods are sold to the consumer and the second term represents the marginal revenue from selling private data. This is the "full" price we normalize to 1.

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

growth rate η such that the optimality conditions for firms and households, the free-entry conditions for innovators, and market clearing conditions hold.

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 for the private sector alone and those that produce for both the government and the private sector.[32] Proposition 2 in Appendix A.2 lays out sufficient conditions for such a BGP to exist and be unique.

We now 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:

$$\eta = \frac{\dot{N}_i}{N_i}. \tag{25}$$

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

$$n_p = \frac{N_p}{N_z}, \tag{26}$$

where N_p is an output-weighted average of private software varieties $N_p \equiv \tilde{N}_p\omega + N_g(1-\omega)$, with $\omega = \frac{q_p(0,p_p,d_p)^{1-\frac{1}{\chi}}}{q_p(0,p_p,d_p)^{1-\frac{1}{\chi}} + q_p(\bar{d}_g,p_p,d_p)^{1-\frac{1}{\chi}}}$.

6 Directed technical change and the role of the state

In this section, we analyze the role of the state in shaping the direction and rate of data-intensive innovation in three applications. We do so by considering alternative state policies that capture three dimensions along which states may vary: first, the development of autocratic surveillance states that monitor and control citizens; second, laws and norms regarding the protection of privacy and the repugnance of markets for data; and third, the state's engagement in industrial policy to promote data-intensive innovation. For each application, we theoretically show how government data non-rivalry and state policies interact to shape macroeconomic outcomes. We then use our reduced-form evidence on data non-rivalry to discipline the model and perform quantitative exercises of interest.

We fully acknowledge the speculative nature of both the theoretical and quantitative results. After all, data-intensive technologies are not yet widespread: their macroeconomic consequences are currently unobservable, and several parameters in the model are hard to

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

pin down with certainty. Yet, we find these results valuable for two reasons. First, the theoretical results serve to illustrate and clarify the similar economic mechanisms underlying the role of the state in all three applications. Second, the quantitative exercises allow us to extrapolate from micro evidence on non-rivalry to the macroeconomic implications of the state's involvement in economies where data-intensive technologies may indeed become widespread, thus providing a sense of the potential impacts of alternative state policies.

6.1 Data-intensive innovation in surveillance states: AI-tocracy

The first application regards innovation under autocratic regimes. A quintessential feature of autocracies is their 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 — by more autocratic states. 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" (see also Zuboff, 2019).[33] China is a prototypical example of this phenomenon: there will be around 560 million public surveillance cameras installed in China by 2021, while approximately 85 million were installed in the US.[34]

Our model and empirical results suggest an alignment between the monitoring aims of autocracies and the innovative aims of AI firms. Consider an extension of our model where the flow utility is:

$$u(C) + \delta G, \tag{27}$$

which captures the social welfare function of a state that values both household utility (either as a benevolent social planner, or selfishly, as the probability of sustaining political power increases with household utility) and also G directly (e.g., an autocracy wanting higher G to better monitor the population and sustain political power).[35] The following theorem then considers the macroeconomic consequences of a larger share of government surveillance spending in total output (i.e., a larger $p_g G/Y$).

Theorem 1 (Rate and direction of innovation in an AI-tocracy) *Assume $\epsilon \geq 1$ and the*

[33]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.

[34]Source: IHS Markit Technology Report, described by the *Wall Street Journal*, `https://bit.ly/33XzsWI`, last accessed on March 30, 2020.

[35]Alternative, this could be interpreted as the utility function of a representative household that cares about private consumption and the security provided by the government. This motive might also be present, for example, in democracies facing security threats.

sufficient conditions in Proposition 2 for a unique BGP equilibrium to exist hold. An economy with a higher share of government surveillance spending in total output ($p_g G/Y$) has a higher rate of innovation (η) and bias of private innovation towards software (n_p) because the supply of government data $\bar{d}_g$ to firms is larger.

Proof. See Appendix A.3. ■

The theorem shows that autocracies' differential demand for monitoring, and thus their differential supply of data to firms, will bias technical change toward data-intensive innovation. Thus, in a balanced growth path of our model, more autocratic states may have *more* rapid technical change in private data-intensive software than less autocratic states. Moreover, the more abundant data in autocratic states will result in higher growth *overall* as well.

We provide an intuitive discussion of the theorem in the Appendix B.1. In brief, the higher profits earned by firms exploiting government data will drive up the interest rate and the rate of innovation on the BGP. In equilibrium, innovators must be indifferent among developing software varieties exploiting government data, developing private software without exploiting government data, and developing non-software varieties. The necessary price adjustments imply that private software will sell at lower prices in the new equilibrium. If relative demand is sufficiently elastic, this implies that the new entry of private software producers will be sufficient to bias technical change toward private software.

Quantitative exercises To get a sense of the magnitude of autocratic advantage in data-intensive innovation, we next calibrate our model. The key challenge is that we have a large degree of uncertainty about the parameters associated with data as an input in innovation: the elasticity of private data supply ψ, 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. We set α to match the value of σ estimated in Section 4.2: the elasticity of private to government software is 0.6, which implies $\sigma = 0.83$. We perform sensitivity analysis for all of these parameters in the Appendix.[36]

We consider the following thought experiment. Imagine the economy associated with our benchmark calibration was the US (assuming a growth rate of 2% per year, the historical average). Holding all else fixed, what are the consequences of increasing government surveillance spending to the levels observed in China? Specifically, we consider a 65 percent increase, which corresponds to the increase from US government spending on domes-

[36] Also see the Appendix C for details on the calibration of the model.

tic security (0.8 percent of GDP) to China's (1.32 percent of GDP) in 2018.[37] Table 4, Panel A.1, shows the results for our benchmark parameterization as well as alternative ones associated with the lower and higher end of our estimates of data non-rivalry. Quantitatively, our parameterized model implies that, compared to a low-surveillance democracy like the US, a surveillance-oriented autocracy like China, which collects and shares more government data as a result of its surveillance demand, would increase its growth rate between 0.6 and 0.95 percentage points and increase the bias of innovation towards private AI software by between 7.8 and 20.9 percent.

Having highlighted an autocratic innovation advantage, we must engage with a previous literature, which has pointed out that autocratic regimes also impose a "tax" on private innovation through the hold up or expropriation of entrepreneurs. It is straightforward to extend our model to incorporate such tax. One way to operationalize it is as a tax on the profits of all innovators selling to the private sector (both software and non-software). This "autocracy tax" has two effects. The first is a direct negative effect on the growth rate, and the second is an indirect effect that increases the relative profitability of selling and innovating on government software. The net effect of the tax on growth is negative.

We next ask: what would the autocracy tax in China need to be such that it features the same growth rate than a democracy like the US but still spends 65 percent more on government surveillance? Table 4, Panel A.2, shows that in the benchmark specification, the "autocracy tax" would need to be 11.5% to completely neutralize the "autocracy advantage" in the data-intensive sector. This is the difference between the corporate tax rate in the notorious tax haven of Ireland, and the tax rate in the United States (Tørsløv et al., 2019). One can also see in Table 4, Panel A.2, that the growth-neutralizing autocracy tax is higher the higher the degree of data non-rivalry: data-intensive surveillance states may impose significant taxes on innovation, yet still grow more quickly than democracies.

6.2 Government data regulation due to privacy concerns

The second application concerns the regulation of government data. The allocation and security of property rights over data are highly variable across countries, reflecting differing legal restrictions on data collection, transmission, and sale, as well as differing norms of privacy, which themselves may influence law. Personal data in particular have an element of "repugnance" (Roth, 2007): many individuals are uncomfortable with their private data being collected and commoditized, in particular when the state is concerned.

[37] 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, Morey et al. (2015) provide interesting evidence that there exists large differences in consumers' willingness to pay to protect their data across the US, UK, Germany, China, and India.[38] On one end of the spectrum, Chinese consumers show low levels of privacy concerns. On the other end of the spectrum, Germans exhibit significant privacy concerns regarding a broad range of data, reflected in, for example, the EU's strict General Data Protection Regulation.[39]

We focus here on norms of privacy that potentially restrict governments' sharing and collection of data. Our model suggests that the expression of local norms regarding data's repugnance can significantly affect the rate and direction of technical change. To see this, consider first a simple extension where the flow utility is:

$$u(C) - \phi_g D_g \tag{28}$$

This captures households' exhibit repugnance towards government data production and data sharing. If households can enforce their preferences through regulations and limits on data collection, production, or sharing, they will be able to reduce D_g in a balanced growth path equilibrium.

The following theorem shows that regulations that force the government to collect and share less data (i.e., a lower D_g/Y) will both the reduce economy's growth rate and bias private sector innovation away from data-intensive goods. The intuition behind this result is analogous to that of lower surveillance spending $p_g G/Y$ in Theorem 1.

Theorem 2 (Data regulation) *Assume $\epsilon \geq 1$ and the sufficient conditions in Proposition 2 for a unique BGP equilibrium to exist hold. A decrease in the share of government data in total output (D_g/Y) will lower the rate of innovation (η) and bias private innovation towards software (n_p decreases).*

Proof. See Appendix A.4. ■

Quantitative exercises To have a sense of magnitudes involved, we use our calibrated model to engage in the following thought experiment. Imagine the economy associated

[38] 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.

[39] 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).

with our benchmark calibration was the US. What are the consequences of decreasing the amount of government data collected and shared to the levels observed in Germany? Specifically, we consider decreasing D_g/Y by 59 percent, which corresponds to the decrease from the number of US cameras (15.28 per 100 residents) to Germany's (6.27 per 100 residents) in 2018.[40] Table 4, Panel B.1, shows the results from this exercise. The growth rate would decrease between 0.4 and 0.9 percentage points — depending on the degree of non-rivalry — and the bias of innovation towards private AI software would decrease between 5 and 22.7 percent.

To conclude this application, we consider a different extension of our benchmark model that further speaks to the ability of the government to collect data. In our benchmark model, government data is produced as a by-product of surveillance alone (see equation 17). Implicitly, this assumes that private transactions (as measured by Y) cannot be tapped to produce government data — either because of technological reasons or, more relevant here, due to regulation that prevents governments from doing so. We now relax this — allowing the government to collect *more* data — and let the government data supply be:

$$D_g = \kappa_g G + \vartheta Y \tag{29}$$

where $\vartheta > 0$ governs the extent to which private transactions can generate government data as a by-product.

To make this new economy comparable to our benchmark one with $\vartheta = 0$, we consider a $\vartheta = 0.0007$ such that the state policy on surveillance spending $p_g G/Y$ remains fixed but government data D_g/Y doubles (which is about the same order of magnitude as the 59 percent decline from our thought experiment above). This requires a 85% increase in $\bar{d}_g$ and a decrease in p_g to occur endogenously.

Table 4, Panel B.2, shows the results. Compared to our benchmark model and parameterization, the economy's growth rate would increase between 0.4 and 0.9 percentage points. The bias of innovation towards private AI software would typically decrease, unless the degree of non-rivalry is sufficiently high.

6.3 Data-driven industrial policy

The third and final application pertains to industrial policy in the age of data-intensive innovation. 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 direction of innovation (and relative size of sectors) as well as increase the overall

[40] 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.

economy's growth rate.

Even without intending to do so, states' provision or restriction of data to private firms may play an important role in driving data-intensive innovation. This implies that industrial policy in the age of data-intensive innovation can also potentially take very different forms.[41] We observed the consequences of the state's public security policies for commercial innovation, due to the non-rivalry of government data, in our empirical analysis in Section 4, and this process is not limited to the Chinese context.[42] Relaxing restrictions on data collection, or reallocating and redefining property rights over data, whether government or private, can also generate large private innovation returns in the data-intensive sector. Public policies like the General Data Protection Regulation in the European Union and the Social Credit System in China will have important industrial policy components (in opposite directions) despite being pursued with very different objectives.[43]

Our theoretical analysis allows us to evaluate the impacts of various forms of (intentional or unintentional) industrial policy. As the next theorem shows, industrial policies that directly provide government data to the private sector *can* achieve the stated goal of increasing the economy's growth rate and biasing the direction of innovation towards private software. However, subsidizing private data production or purchases alone is *less* likely to achieve such goals. In fact, in our benchmark model, both the rate and direction of private innovation turn out to be neutral with respect to such private data subsidies along the BGP. We next state the theorem and then discuss which assumptions need to be relaxed in other to break the neutrality result.[44]

Theorem 3 (Industrial policy) *Assume $\epsilon \geq 1$ and the sufficient conditions in Proposition 2 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 software (increase n_p). A subsidy to private data production (a decrease in τ) will increase the share of private data in total*

[41] This echoes the argument of Rodrik (2007) that all policies, whether intended or not, can be considered as industrial policies. For example, deregulating banking has the incidental industrial policy consequence of helping more capital-intensive sectors.

[42] 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.

[43] The former is an attempt 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. The latter is an attempt to control the Chinese population by collecting and linking multidimensional personal data, again with important consequences for innovation.

[44] Note also that the effects of changes in private data subsidies are equivalent, albeit in the opposite direction, to changes in regulations that effectively "tax" private data collection and sharing. Such regulations could result from privacy considerations and repugnance as in our previous application.

output (D_p/Y) when its supply is elastic ($\psi < 1$). However, the subsidy will not affect η or n_p.

Proof. See Appendix A.5. ■

The consequences of a change in $\bar{d}_g$ are analogous to those arising from a change in surveillance spending $p_g G/Y$ in Theorem 1, or government data as a share of total output in Theorem 2. Regarding the neutrality of private data subsidies in our benchmark economy, this result is surprising in light of the typical view that industrial policies that subsidize key inputs can affect macroeconomic outcomes.[45] As it turns out, the result is driven by one important difference between typical economies and the data-intensive economy we study in our benchmark model: the economics of government data. Specifically, the state is a supplier of government data, which is a substitute to the private data being subsidized, and the production of government data is itself a by-product of *only* government-data-intensive surveillance. If the state does not demand surveillance or supply data ($G = \bar{d}_g = 0$) or if government data is a by-product of *both* surveillance and private transactions (for example, $D_g = \kappa_g G + \vartheta Y$, as in our extension from the previous application), then decreases in τ will increase the growth rate (η) and bias private innovation towards software (increase n_p).[46] We next quantitatively study policies affecting the availability of government data as well as traditional industrial policy.

Quantitative exercises We first consider increases in government data made available for firms that are sufficiently large to induce an increase in the economy's growth rate by 0.1% across BGP equilibria. Second, we consider subsidies to private data production that induce the same growth rate increase. In both cases, we calculate the fiscal cost as measured by the percentage change in total government expenditures as a fraction of total output: $-\tau p_d D_p/Y + p_g G/Y$. Table 4, Panels C.1 and C.2 report the results for different combinations of ϑ and ψ — which govern the fraction of total government data that is produced as a by-product of private transactions, and the elasticity of private data supply, respectively.

On the one hand, we find that increases in government data are typically effective (and have a similar fiscal cost) regardless of ϑ and ψ. On the other hand, the effect of subsidies on private data are highly sensitive to such features of the economy. For example, when most government data is produced as a by-product of government surveillance ($\vartheta \to 0$), then increasing the economy's growth rate via private data subsidies becomes fiscally prohibitive. This is simply a reflection of the neutrality result in Theorem 3. Moreover, even

[45] We provide the intuition behind Theorem 3 in the Appendix B.2.

[46] When $G = \bar{d}_g = 0$, our model reduces to a standard two-sector directed-technical change model where the common view of industrial policy holds. The key difference relative to the benchmark model is that the demand for private data (the key input subsidized) is no longer infinitely elastic in a BGP. When $D_g = \kappa_g G + \vartheta Y$, again private demand demand is not infinitely elastic in a BGP.

when ϑ is larger, subsidies on private data become fiscally more attractive than increasing government data only if private data supply is sufficiently elastic (low ψ). Given our parameterization, the increase in the fiscal cost of both policies is exactly identical (13%) when $\psi = 0.85$. Whenever ψ is lower than that, then subsidies become more attractive. For instance, when $\psi = 0.6$, the increase in fiscal costs are 3.8% and 17% respectively.

7 Conclusion

Taken together, our analysis sheds new light on the role of the state in data-intensive economies. First, traditional models of growth under autocracy emphasize an implicit "tax" on innovation. Through the lens of such models, it is difficult to explain how Chinese research exploded to overtake the US as the largest source of highly-cited AI work (see Appendix Figure A.16), or Chinese firms' global leadership in AI innovation.[47] Our work highlights an autocratic advantage in data-intensive innovation arising from the non-rivalry of data. In our framework, one can explain China's rise to pre-eminence in AI innovation not *despite* its political institutions, but in part *because* of them.

Second, while traditional models of property rights might suggest restrictions on the collection and sharing of individual data, our model suggests that such restrictions can hinder innovation and growth when individual data are an important input in the innovation process.

Third, while traditional models of industrial policy suggest that subsidies lowering prices in private markets can have stimulating effects at the sectoral or aggregate level, our model suggests that such subsidies might be prohibitively expensive when government data plays an important role in generating innovation. "Unintentional" industrial policy — working through the state's provision of government data to private firms — might be more effective.

We thus conclude that states — through their collection, distribution, and regulation of data — will be critical in shaping the global landscape in the coming decades as data-intensive technologies become widespread.

[47] Consider the facial recognition AI sector as an example: Appendix Figure A.17 presents the 2018 ranking of the companies with the top facial recognition algorithms, as ranked by the Face Recognition Vendor Test (FRVT), organized by the NIST, 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 broader 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.

_ **and Pascual Restrepo**, "Artificial Intelligence, Automation and Work," *NBER Working Paper*, January 2018, pp. 1–43.

_ **, 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.

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.

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

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

Bessen, James, "Ai and Jobs: the Role of Demand," *NBER Working Paper*, January 2018, pp. 1–27.

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.

Cockburn, Iain M, Rebecca Henderson, and Scott Stern, "The Impact of Artificial Intelligence on Innovation," *NBER Working Paper*, March 2018, pp. 1–40.

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.

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.

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.

Korinek, Anton and Joseph E Stiglitz, "Artificial Intelligence and Its Implications for Income Distribution and Unemployment," *NBER Working Paper*, December 2017, pp. 1–45.

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.

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.

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.

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

Sachs, Jeffrey D, "R&D, Structural Transformation, and the Distribution of Income," in "The Economics of Artificial Intelligence An Agenda" January 2018, pp. 1–25.

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

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.

Stevenson, Betsey, "Artificial Intelligence, Income, Employment, and Meaning," in Ajay Agrawal, Joshua Gans, and Avi Goldfarb, eds., *The Economics of Artificial Intelligence An Agenda*, University of Chicago Press, 2019, pp. 1–1.

Tørsløv, Thomas, Ludvig Wier, and Gabriel Zucman, "The Missing Profits of Nations," *Working Paper*, September 2019, pp. 1–48.

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

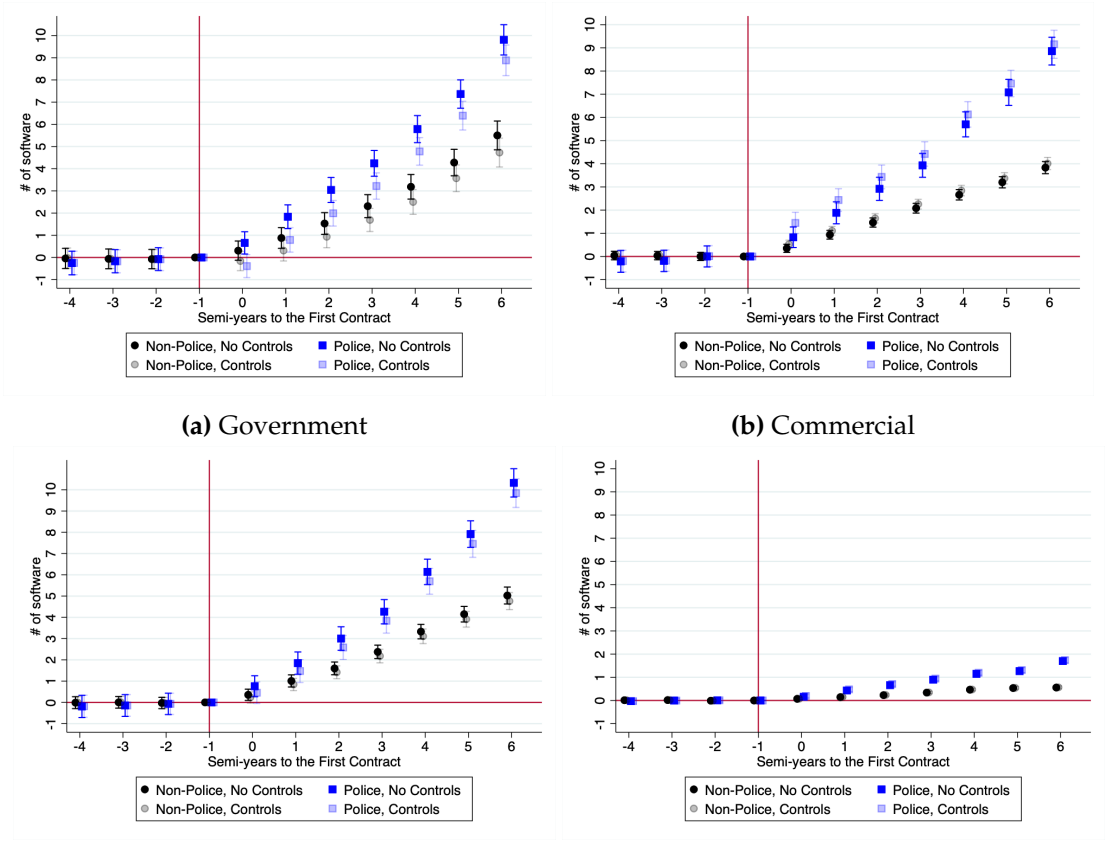
Varian, Hal R, "Artificial Intelligence, Economics, and Industrial Organization," *NBER Working Paper*, July 2018, pp. 1–24.

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.

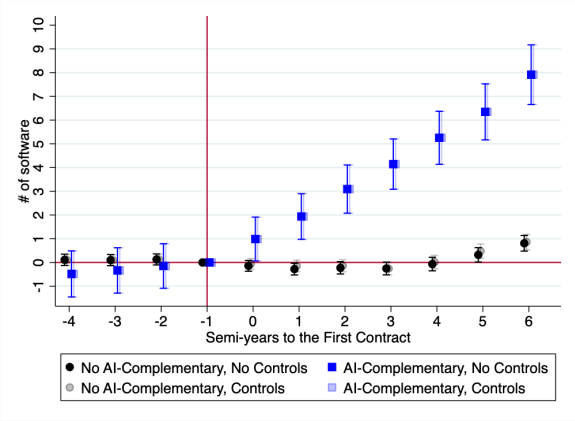
Figures and tables



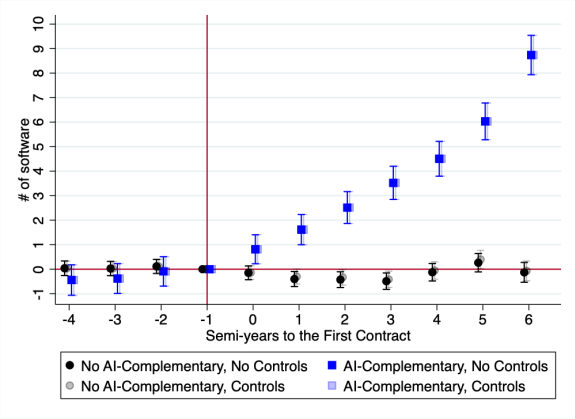
(a) Government **(b)** Commercial

(c) Data-Complementary **(d)** Video-Related AI

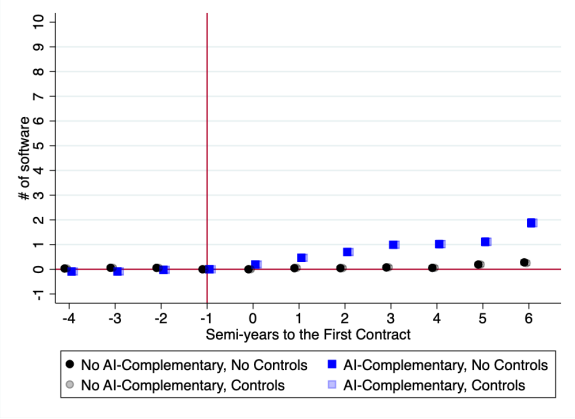
Figure 1: Software development intended for government (top panel – left), and for commercial uses (top panel – right), relative to the time of receiving initial procurement contract, controlling for firms and time period fixed effects. Data-complementary software development — data storage, data transmission, or data management (bottom panel – left), and video-related AI software (bottom panel – right), relative to the time of receiving initial procurement contract. Blue dots are police contracts, black non-police. Translucent lines/markers additionally interacts outcomes with pre-period variables including area gdp, firm size, and contract size.



(a) Government



(b) Commercial



(c) Video-Related AI

Figure 2: Relative increase in software production between receiving the first public security versus non-public security contracts, separately estimated for firms that have produced at least one data-complementary software (data storage, data transmission, or data management) during the first 12 months since the intial contract was signed (blue dots), and those that have not produced any data-complementary software (black squares). Figures correspond to different software: software development intended for government (top panel), intended for commercial use (middle panel), and video-related AI software (bottom panel). Translucent lines/markers additionally interacts outcomes with pre-period variables including area gdp, firm size, and contract size. All estimates control for firms and time period fixed effects.

Table 1: Summary stats table for contracts: procurement contracts

	Public security contracts		Non-public security contracts	
	All	1st	All	1st
	(1)	(2)	(3)	(4)
Monetary size (10,000 RMB)	3,993,643.67	2,596,616.23	2,883,409.70	1,410,223.44
	(37,781,759.13)	(13,598,359.41)	(37,762,942.85)	(12,556,273.17)
Provincial or above admin. level	0.28	0.23	0.34	0.45
	(0.45)	(0.42)	(1.01)	(0.50)
Years when contracts signed	2,016.38	2,015.74	2016.22	2015.9
	(1.61)	(1.93)	(1.61)	(1.84)
Observations	16,483	569	12,309	986

Note: The analysis unit is procurement contract. 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.

Table 2: Summary statistics — firms and their productions

		Firms with public security contracts			Firms without
	All firms	All	High capacity	Low capacity	public security contracts
	(1)	(2)	(3)	(4)	(5)
Panel A: firm characteristics					
Years of establishment	2,013.26	2,008.14	2007.98	2008.30	2,013.66
	(4.85)	(6.83)	(6.74)	(6.91)	(4.42)
Capital (Million US$)	7.32	32.67	24.80	40.31	5.33
	(87.68)	(256.20)	(116.71)	(341.05)	(55.57)
Total # of investment fundings	2.06	2.74	2.49	2.71	2.01
	(2.93)	(1.96)	(1.48)	(1.78)	(2.98)
Panel B: software production					
Total # of software	13.47	44.28	45.24	43.35	10.51
	(29.44)	(80.47)	(86.06)	(74.87)	(15.12)
Government	5.96	18.75	19.24	18.76	4.30
	(14.71)	(38.32)	(37.00)	(38.73)	(5.48)
Commercial	6.59	16.42	16.60	16.25	5.50
	(13.70)	(34.92)	(38.72)	(30.85)	(7.90)
AI-Common	3.97	9.39	9.80	9.36	3.26
	(7.47)	(17.22)	(17.64)	(16.86)	(4.47)
AI-Video	2.68	4.98	4.86	5.08	2.26
	(3.35)	(6.18)	(6.33)	(6.06)	(2.26)
AI-Complementary	6.77	19.54	20.26	18.81	5.28
	(14.24)	(36.56)	(41.65)	(30.73)	(7.07)
Panel C: patent development					
Total # of patents	51.70	175.85	179.72	172.16	39.20
	(284.71)	(623.85)	(653.98)	(595.19)	(220.05)
Government	15.85	44.77	48.41	41.40	12.17
	(73.17)	(140.02)	(164.38)	(113.35)	(58.53)
Commercial	27.25	75.51	82.54	68.98	21.98
	(160.39)	(304.27)	(361.93)	(239.34)	(134.83)
AI-Common	14.48	40.31	43.45	37.66	11.34
	(66.50)	(138.32)	(142.50)	(135.04)	(50.47)
AI-Video	10.10	32.28	28.92	35.32	7.15
	(40.63)	(94.06)	(87.58)	(99.70)	(24.90)
AI-Complementary	23.62	77.45	85.39	70.17	17.45
	(120.89)	(263.70)	(297.71)	(228.63)	(89.21)
Observations	7837	568	282	287	7269

Note: The unit of the summary statistics is subsidiary firm. Standard deviations are reported below the means.

Table 3: Effects of public security contracts on software development

	Government	Commercial	AI-complementary	AI-video	Government	Commercial	AI-complementary	AI-video
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
4 Semiyears Before	-0.112	0.004	-0.114	0.000	-0.016	0.019	-0.014	0.008
	(0.196)	(0.138)	(0.157)	(0.026)	(0.218)	(0.148)	(0.175)	(0.027)
3 Semiyears Before	-0.114	-0.004	-0.071	0.005	-0.055	0.011	-0.009	0.012
	(0.192)	(0.136)	(0.154)	(0.026)	(0.214)	(0.145)	(0.171)	(0.027)
2 Semiyears Before	-0.103	-0.033	-0.075	-0.022	-0.082	-0.006	-0.038	-0.013
	(0.187)	(0.133)	(0.151)	(0.025)	(0.209)	(0.142)	(0.168)	(0.026)
Receiving 1st Contract	0.396**	0.397***	0.438***	0.072***	-0.122	0.518***	0.239	0.078***
	(0.187)	(0.132)	(0.150)	(0.025)	(0.209)	(0.142)	(0.168)	(0.026)
1 Semiyear After	0.858***	0.883***	0.990***	0.124***	0.226	0.968***	0.744***	0.127***
	(0.201)	(0.142)	(0.161)	(0.027)	(0.225)	(0.152)	(0.179)	(0.028)
2 Semiyears After	1.410***	1.317***	1.510***	0.231***	0.798***	1.514***	1.233***	0.230***
	(0.211)	(0.149)	(0.169)	(0.028)	(0.235)	(0.159)	(0.188)	(0.030)
3 Semiyears After	2.145***	1.971***	2.258***	0.326***	1.595***	2.130***	2.082***	0.346***
	(0.223)	(0.158)	(0.179)	(0.030)	(0.249)	(0.169)	(0.199)	(0.031)
4 Semiyears After	3.067***	2.566***	3.273***	0.455***	2.371***	2.755***	3.056***	0.484***
	(0.238)	(0.168)	(0.192)	(0.032)	(0.265)	(0.180)	(0.213)	(0.034)
5 Semiyears After	3.809***	2.957***	3.925***	0.504***	3.374***	3.287***	3.764***	0.538***
	(0.256)	(0.181)	(0.207)	(0.034)	(0.286)	(0.194)	(0.229)	(0.036)
6 Semiyears After	4.918***	3.754***	4.838***	0.534***	4.675***	4.000***	4.648***	0.582***
	(0.278)	(0.197)	(0.224)	(0.037)	(0.310)	(0.210)	(0.248)	(0.039)
4 Semiyears Before × Public Security	-0.143	-0.133	-0.072	-0.022	-0.147	-0.130	-0.104	-0.024
	(0.317)	(0.225)	(0.255)	(0.042)	(0.354)	(0.240)	(0.284)	(0.045)
3 Semiyears Before × Public Security	-0.047	-0.135	-0.067	-0.006	-0.069	-0.133	-0.090	-0.012
	(0.311)	(0.220)	(0.250)	(0.042)	(0.347)	(0.236)	(0.278)	(0.044)
2 Semiyears Before × Public Security	0.020	0.061	0.012	0.039	0.040	0.047	-0.002	0.020
	(0.306)	(0.217)	(0.246)	(0.041)	(0.342)	(0.232)	(0.274)	(0.043)
Receiving 1st Contract × Public Security	0.422	0.465**	0.280	0.109***	-0.351	0.497**	0.022	0.066
	(0.301)	(0.214)	(0.243)	(0.040)	(0.337)	(0.229)	(0.270)	(0.042)
1 Semiyear After × Public Security	0.874***	0.712***	0.611**	0.247***	0.263	1.005***	0.396	0.211***
	(0.321)	(0.228)	(0.259)	(0.043)	(0.359)	(0.244)	(0.288)	(0.045)
2 Semiyears After × Public Security	1.392***	1.293***	1.261***	0.366***	0.754**	1.337***	1.094***	0.334***
	(0.337)	(0.239)	(0.271)	(0.045)	(0.376)	(0.256)	(0.302)	(0.048)
3 Semiyears After × Public Security	1.807***	1.643***	1.599***	0.490***	1.141***	1.712***	1.408***	0.435***
	(0.351)	(0.249)	(0.283)	(0.047)	(0.392)	(0.267)	(0.314)	(0.050)
4 Semiyears After × Public Security	2.333***	2.681***	2.359***	0.546***	1.937***	2.754***	2.156***	0.507***
	(0.370)	(0.262)	(0.298)	(0.050)	(0.413)	(0.281)	(0.332)	(0.052)
5 Semiyears Before × Public Security	3.068***	3.610***	3.386***	0.634***	2.350***	3.618***	3.173***	0.607***
	(0.392)	(0.277)	(0.315)	(0.052)	(0.437)	(0.296)	(0.350)	(0.055)
6 Semiyears After × Public Security	4.150***	4.360***	4.692***	0.980***	3.496***	4.519***	4.630***	0.947***
	(0.421)	(0.297)	(0.338)	(0.057)	(0.470)	(0.317)	(0.376)	(0.060)
Observations	1.19e+05	1.20e+05	1.20e+05	1.16e+05	1.19e+05	1.20e+05	1.20e+05	1.16e+05

Notes Baseline specification controls for time period fixed effects, and subsidiary firm fixed effects. Standard errors clustered at mother firm level are reported below the coefficients. Columns 5-8 include controls for monetary size of contracts, firms' pre-contract characteristics, and 1st contract's location characteristics (local GDP at the year of contract), and their interaction with pre-1st contract semi-year indicators. * significant at 10% ** significant at 5% *** significant at 1%.

Table 4: The role of the state: a calibration exercise

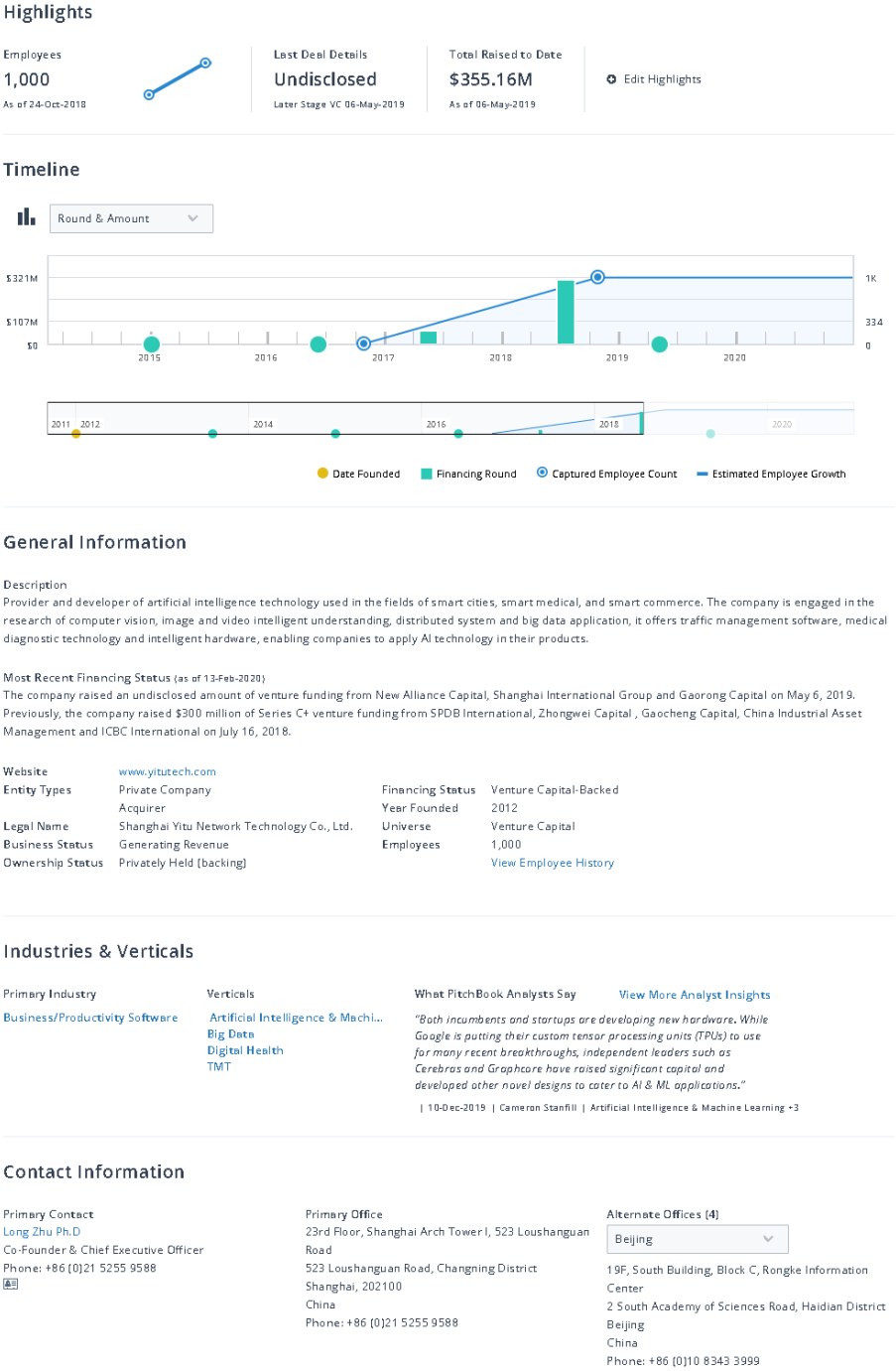
	Benchmark $\sigma = 0.83$	Low $\sigma = 0.68$	High $\sigma = 0.97$
Panel A: innovation under autocracy			
A.1: from US to China's government surveillance spending			
$\Delta log(n_p)$	7.8%	3.9%	20.9%
$\Delta\eta$	0.6%	0.4%	1.0%
A.2: the growth-neutralizing autocracy tax			
Autocracy tax	11.5%	8.5%	17.5%
Panel B: government data regulation			
B.1: from US to Germany's cameras per capita			
$\Delta log(n_p)$	-10.7%	-5.0%	-22.7%
$\Delta\eta$	-0.6%	-0.4%	-0.9%
B.2: allow private transactions in producing govt. data			
$\Delta log(n_p)$	-3.0%	-3.9%	2.5%
$\Delta\eta$	0.5%	0.4%	0.9%
Panel C: data-driven industrial policy			
C.1: allow private transactions in producing govt. data			
$\Delta log(Fiscal_cost)$	11%	13%	17%
C.2: subsidy on private data			
$\Delta log(Fiscal_cost)$	∞	27%	3.8%

Notes: Panel A.1 reports the percentage changes in the rate η and bias of private innovation n_p following a 65 percent increase in the government spending to output ratio. Panel A.2 reports the growth-neutralizing autocracy tax. Panel B.1 reports the percentage changes in the rate η and bias of private innovation n_p following a 59 percent increase in the government data to output ratio. Panel B.2 reports changes following a new policy with that allows private transactions to be used in government data production ($\vartheta > 0$) such that the government data to output doubles but surveillance spending remains the same. 'Fiscal cost' in Panels C.1 and C.2 is defined as total government expenditures as a fraction of total output: $-\tau p_d D_p/Y + p_g G/Y$. We choose $\vartheta > 0$ so that 10 percent of total government data is produced as by-product of private transactions. For each ϑ, ψ combination in each column, we re-calibrate all other parameters so that the initial economy corresponds to our benchmark symmetric economy. Panel C.1 examines increase in government data available ($\Delta\bar{d}_g > 0$) such that $\Delta\eta = 0.1\%$, and reports the percentage change in the fiscal cost that is necessary in order to increase the economy's growth rate by 0.1% via increases in government data available to firms. Panel C.2 repeats the exercise, but instead examines subsidy on private data ($\Delta\tau < 0$) such that $\Delta\eta = 0.1\%$. 'Benchmark' corresponds to benchmark parameterization with $\sigma = 0.8$. 'Low σ' re-parameterizes the model using $\sigma = 0.68$, which corresponds to matching +2SD of the government spending response and -2SD of the private response in Table 3 at a 2 year horizon. 'High σ' re-parameterizes the model using $\sigma = 0.97$, which correspond to matching -2SD of the government spending response and +2SD of the private response in Table 3 at a 2 year horizon.

ONLINE APPENDIX (NOT FOR PUBLICATION)



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



Highlights

Employees
1,000
As of 24-Oct-2018

Last Deal Details
Undisclosed
Later Stage VC 06-May-2019

Total Raised to Date
$355.16M
As of 06-May-2019

Edit Highlights

Timeline

Round & Amount

$321M
$107M
$0
1K
334
0
2015
2016
2017
2018
2019
2020
2011
2012
2014
2016
2018
2020

Date Founded
Financing Round
Captured Employee Count
Estimated Employee Growth

General Information

Description
Provider and developer of artificial intelligence technology used in the fields of smart cities, smart medical, and smart commerce. The company is engaged in the research of computer vision, image and video intelligent understanding, distributed system and big data application, it offers traffic management software, medical diagnostic technology and intelligent hardware, enabling companies to apply AI technology in their products.

Most Recent Financing Status (as of 13-Feb-2020)
The company raised an undisclosed amount of venture funding from New Alliance Capital, Shanghai International Group and Gaorong Capital on May 6, 2019. Previously, the company raised $300 million of Series C+ venture funding from SPDB International, Zhongwei Capital , Gaocheng Capital, China Industrial Asset Management and ICBC International on July 16, 2018.

Website	www.yitutech.com		
Entity Types	Private Company	Financing Status	Venture Capital-Backed
	Acquirer	Year Founded	2012
Legal Name	Shanghai Yitu Network Technology Co., Ltd.	Universe	Venture Capital
Business Status	Generating Revenue	Employees	1,000
Ownership Status	Privately Held (backing)		View Employee History

Industries & Verticals

Primary Industry
Business/Productivity Software

Verticals
Artificial Intelligence & Machi...
Big Data
Digital Health
TMT

What PitchBook Analysts Say
View More Analyst Insights

"Both incumbents and startups are developing new hardware. While Google is putting their custom tensor processing units (TPUs) to use for many recent breakthroughs, independent leaders such as Cerebras and Graphcore have raised significant capital and developed other novel designs to cater to AI & ML applications."

| 10-Dec-2019 | Cameron Stanfill | Artificial Intelligence & Machine Learning +3

Contact Information

Primary Contact
Long Zhu Ph.D
Co-Founder & Chief Executive Officer
Phone: +86 (0)21 5255 9588

Primary Office
23rd Floor, Shanghai Arch Tower I, 523 Loushanguan Road
523 Loushanguan Road, Changning District
Shanghai, 202100
China
Phone: +86 (0)21 5255 9588

Alternate Offices (4)
Beijing

19F, South Building, Block C, Rongke Information Center
2 South Academy of Sciences Road, Haidian District
Beijing
China
Phone: +86 (0)10 8343 3999

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



Figure A.3: Example of a procurement contract record.

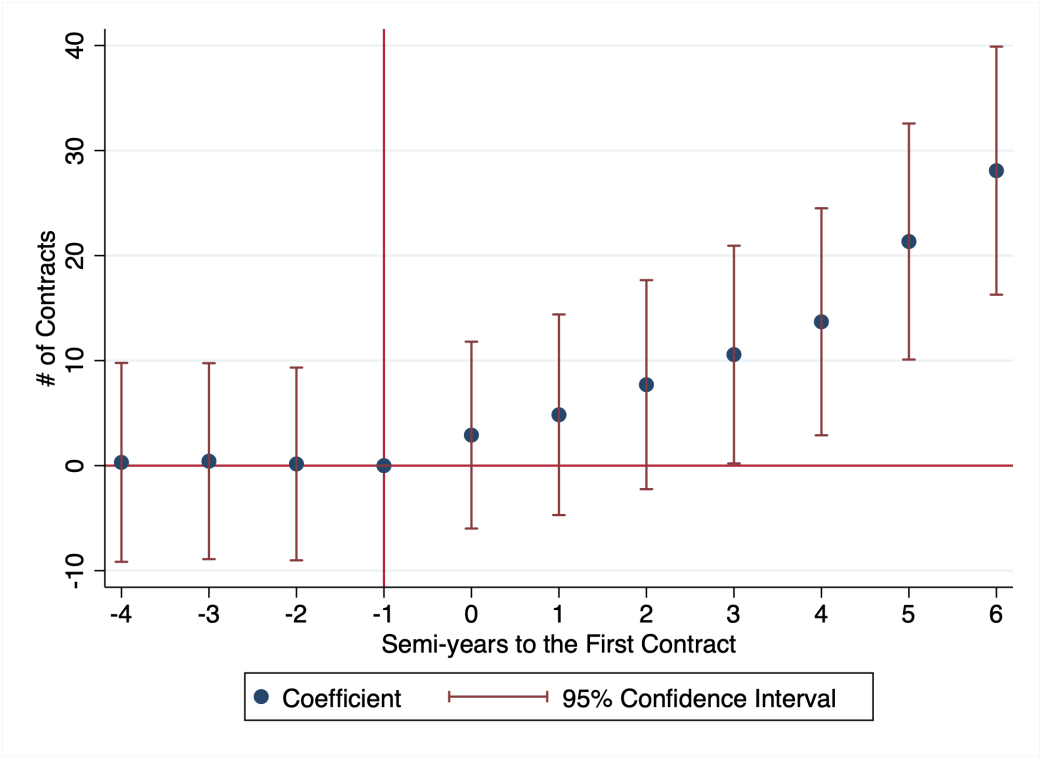


Figure A.4: Arrival of public security procurement contracts after the inital contract arrival.

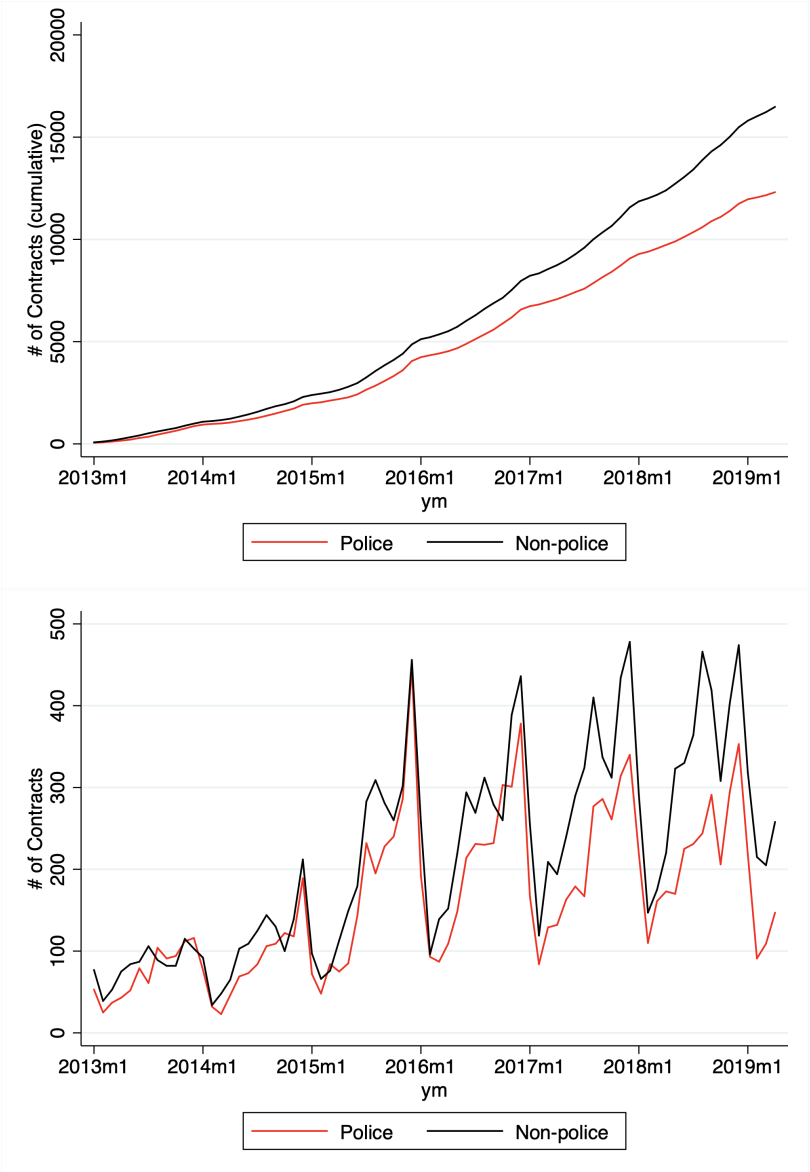
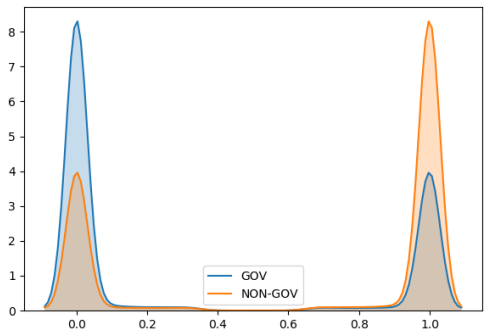
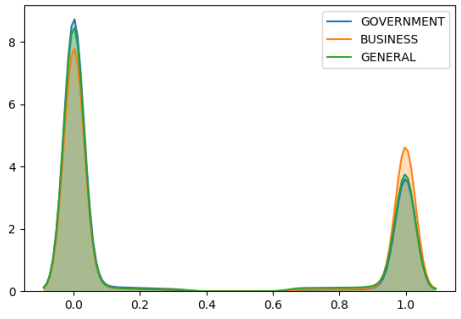


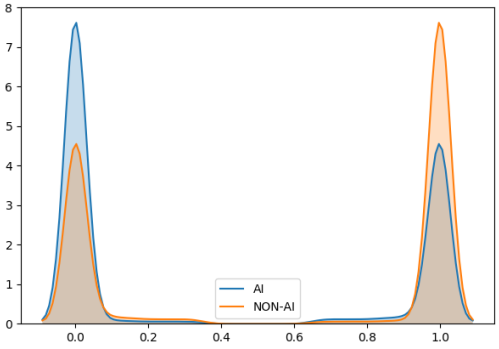
Figure A.5: Cumulative number of public security and non-public security contracts (top panel), and the flow of new contracts signed in each month (bottom panel).



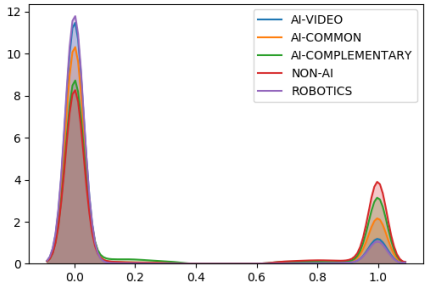
(a) Customers - Binary



(b) Customers - Non-Binary

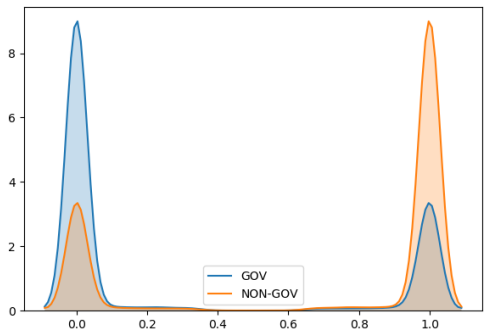


(c) Function - Binary

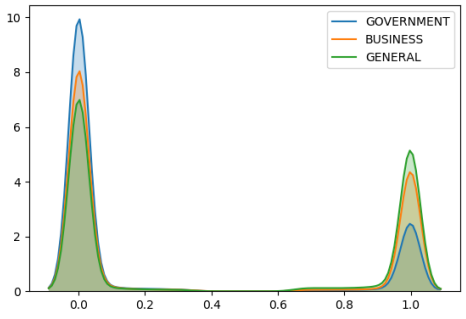


(d) Function - Non-Binary

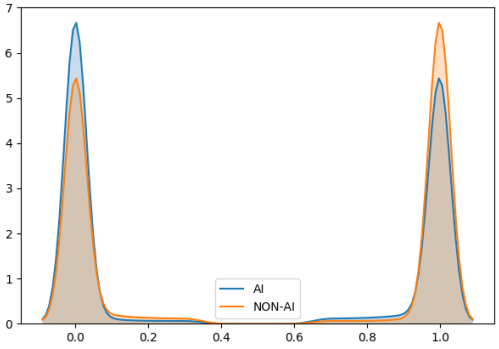
Figure A.6: Probability density plots of software categorizations based on Recurrent Neural Network with Long Short-Term Memory algorithm. Top panel shows categorization on customers (left is binary; right is full set of categories); bottom panel shows categorization on function (left is binary; right is full set of categories).



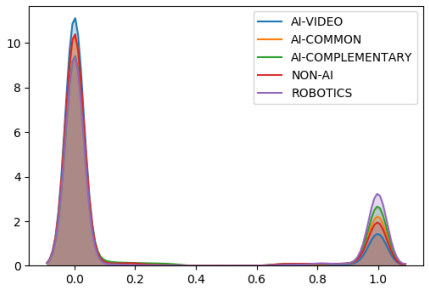
(a) Customers - Binary



(b) Customers - Non-Binary

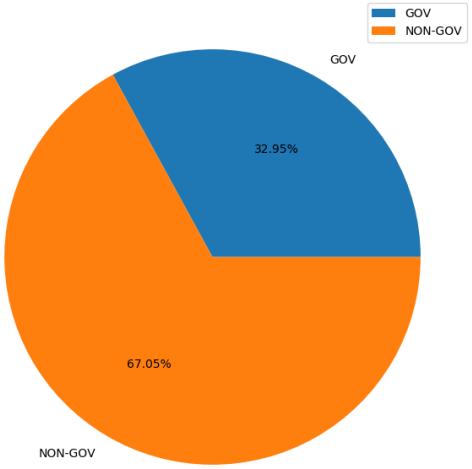


(c) Function - Binary

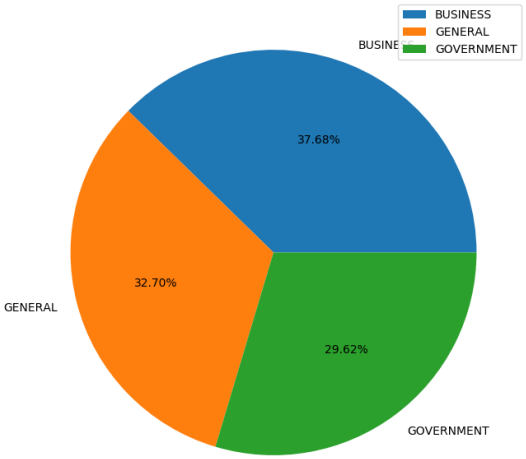


(d) Function - Non-Binary

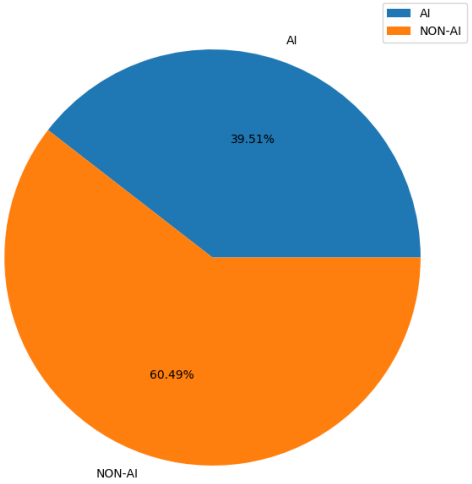
Figure A.7: Probability density plots of patent categorizations based on Recurrent Neural Network with Long Short-Term Memory algorithm. Top panel shows categorization on customers (left is binary; right is full set of categories); bottom panel shows categorization on function (left is binary; right is full set of categories).



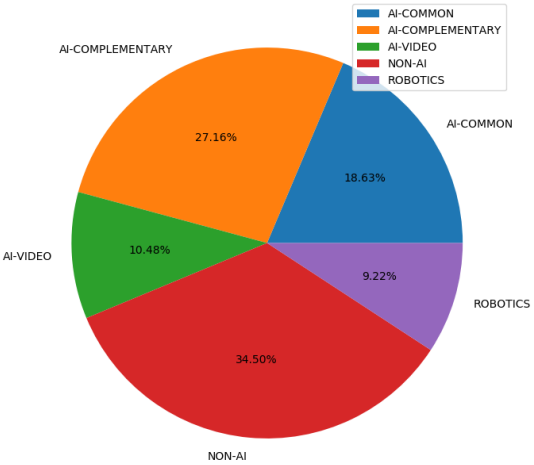
(a) Customers - Binary



(b) Customers - Non-Binary

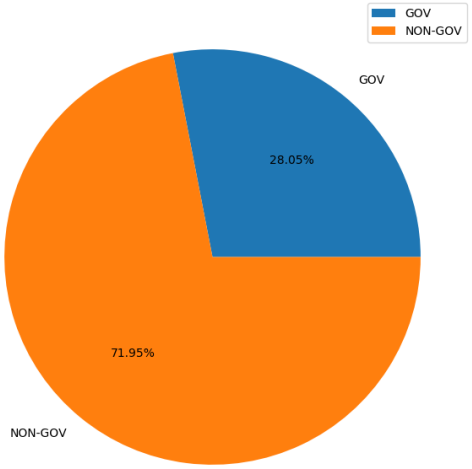


(c) Function - Binary

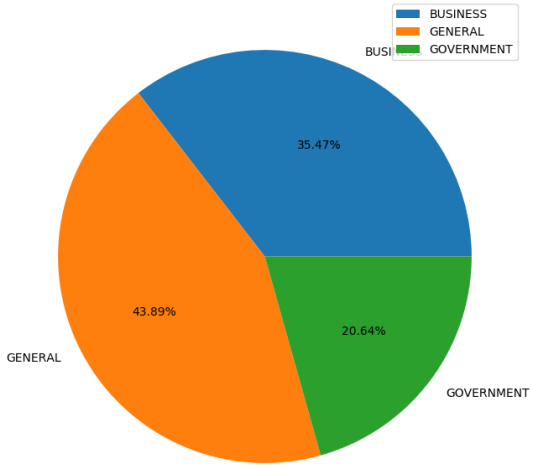


(d) Function - Non-Binary

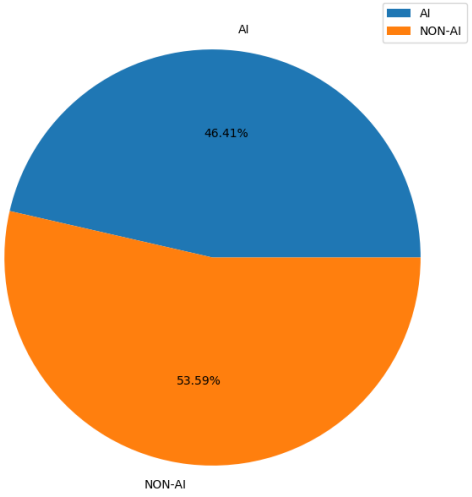
Figure A.8: Summary statistics of categorization outcomes for software categorizations based on Recurrent Neural Network with Long Short-Term Memory algorithm. Top panel shows categorization on customers (left is binary; right is full set of categories); bottom panel shows categorization on function (left is binary; right is full set of categories).



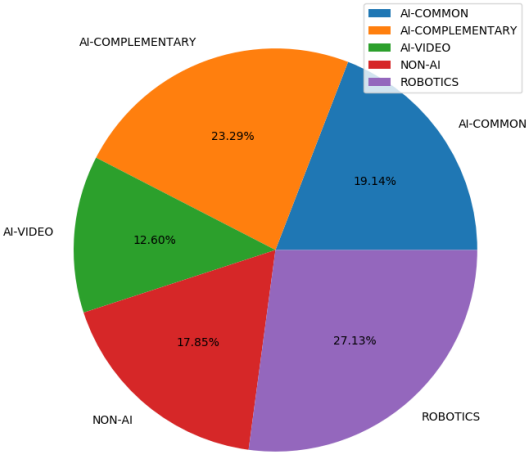
(a) Customers - Binary



(b) Customers - Non-Binary

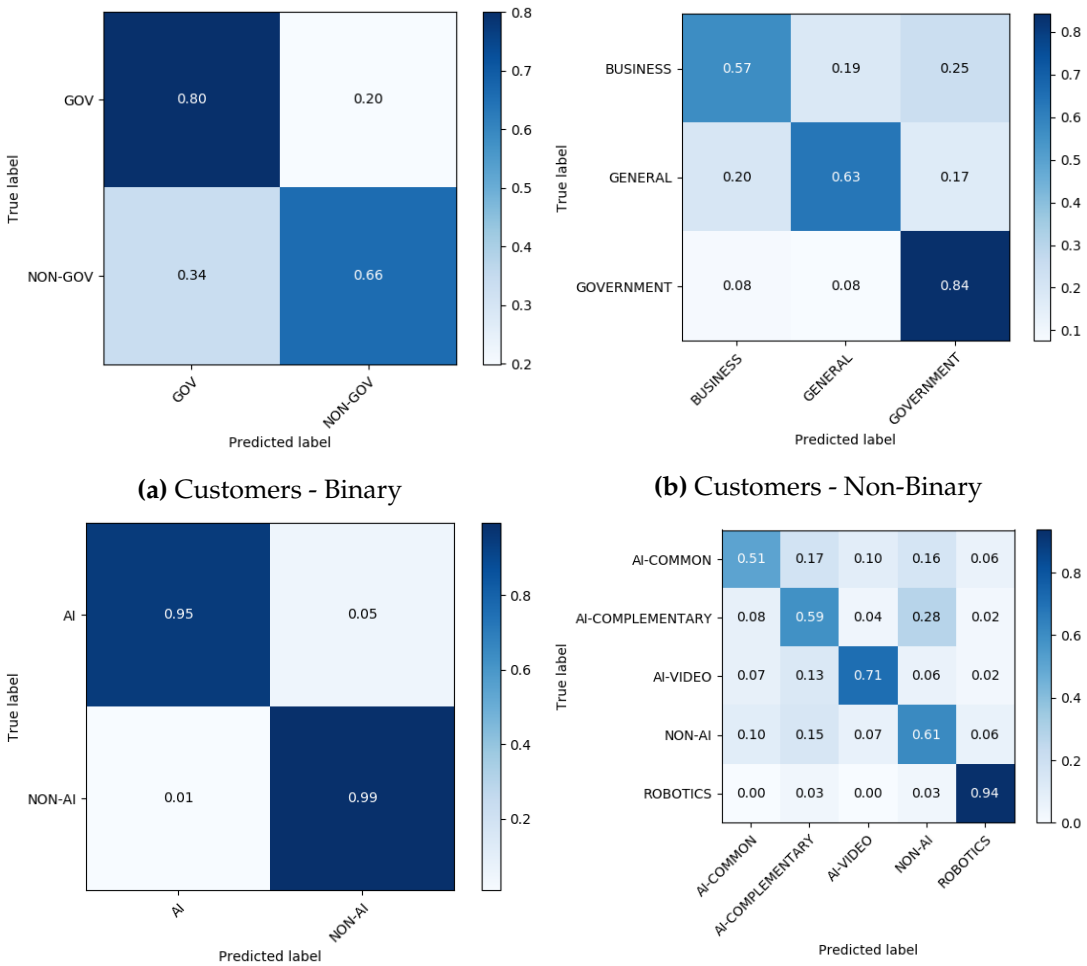


(c) Function - Binary



(d) Function - Non-Binary

Figure A.9: Summary statistics of categorization outcomes for patent categorizations based on Recurrent Neural Network with Long Short-Term Memory algorithm. Top panel shows categorization on customers (left is binary; right is full set of categories); bottom panel shows categorization on function (left is binary; right is full set of categories).



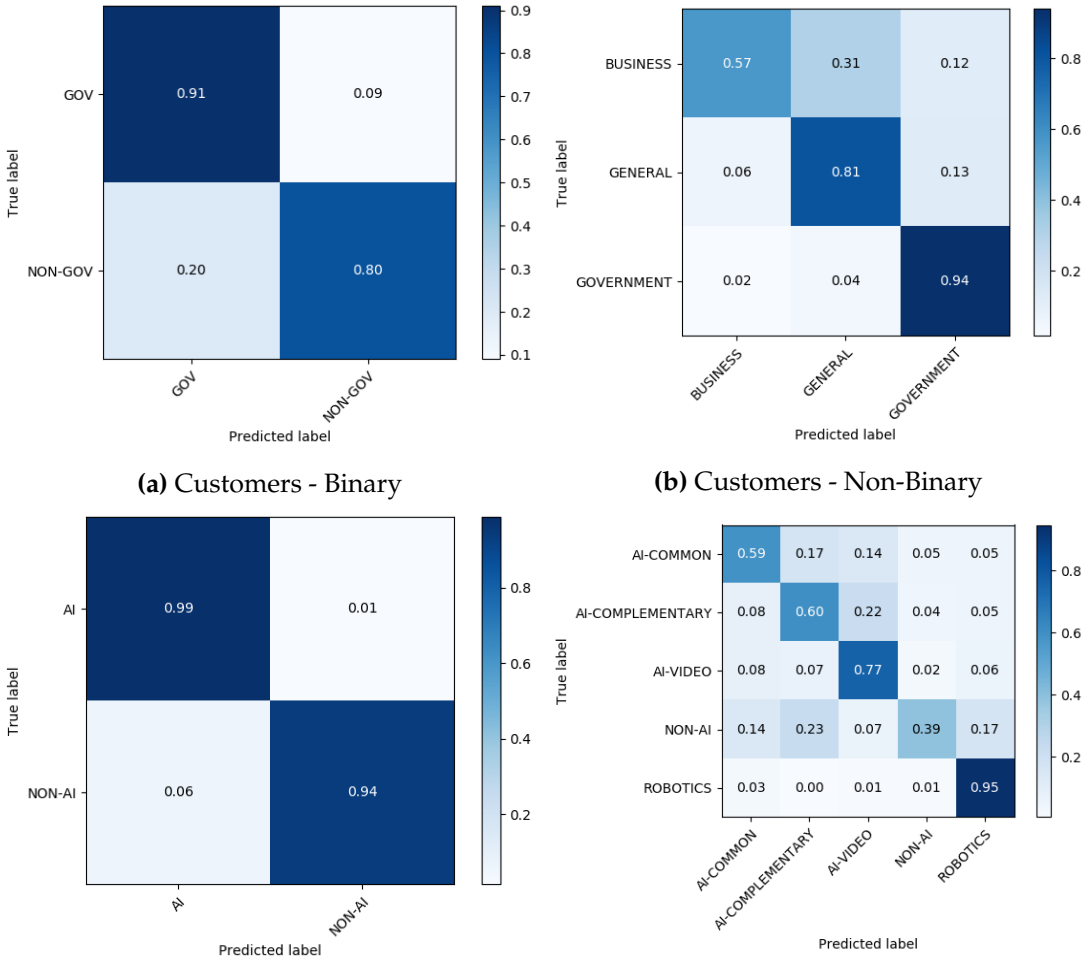
(a) Customers - Binary

(b) Customers - Non-Binary

(c) Functions - Binary

(d) Functions - Non-Binary

Figure A.10: 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 on customers (left is binary; right is full set of categories); bottom panel shows categorization on 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.11: Confusion matrix of categorization outcomes for patent 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 on customers (left is binary; right is full set of categories); bottom panel shows categorization on function (left is binary; right is full set of categories).

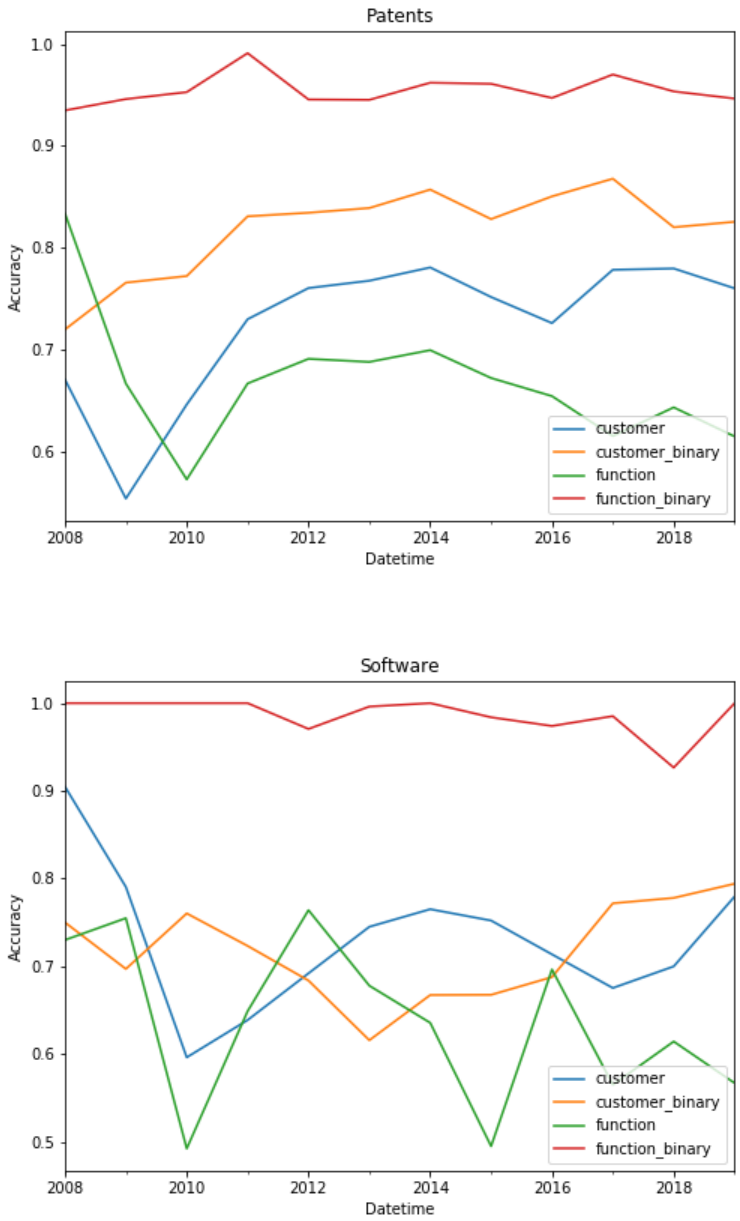
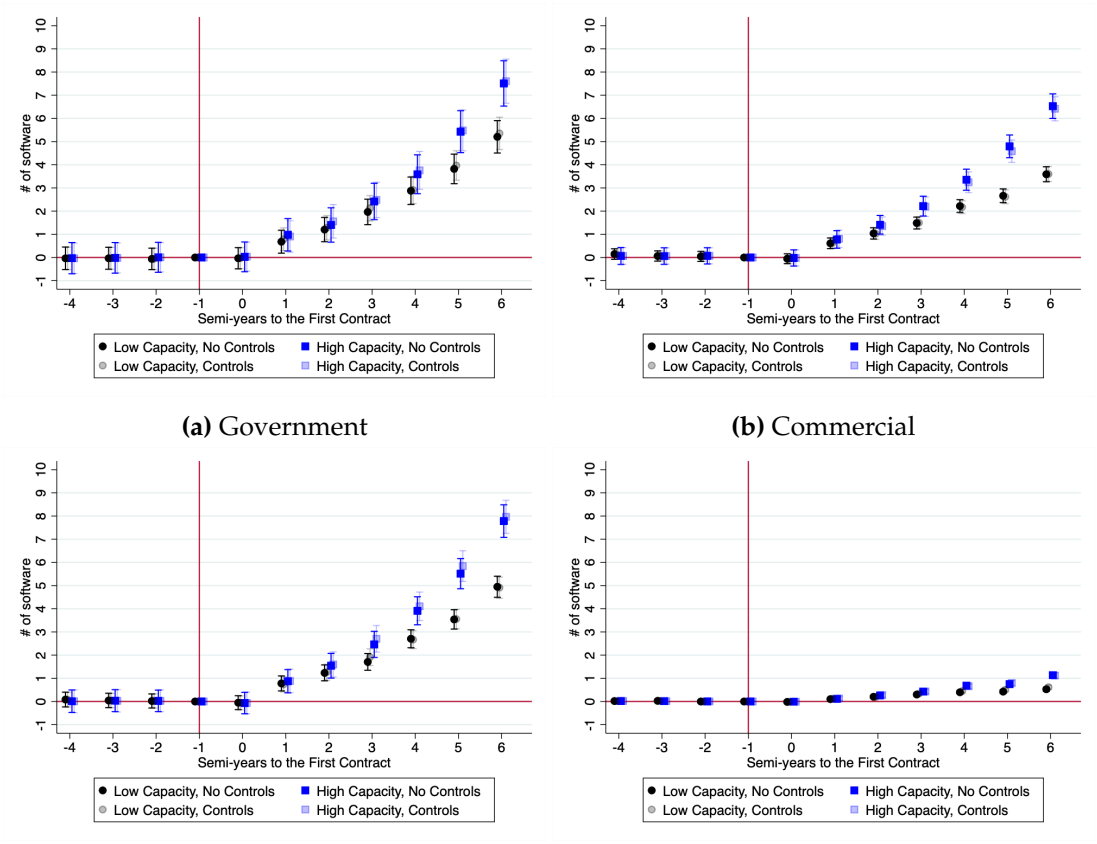


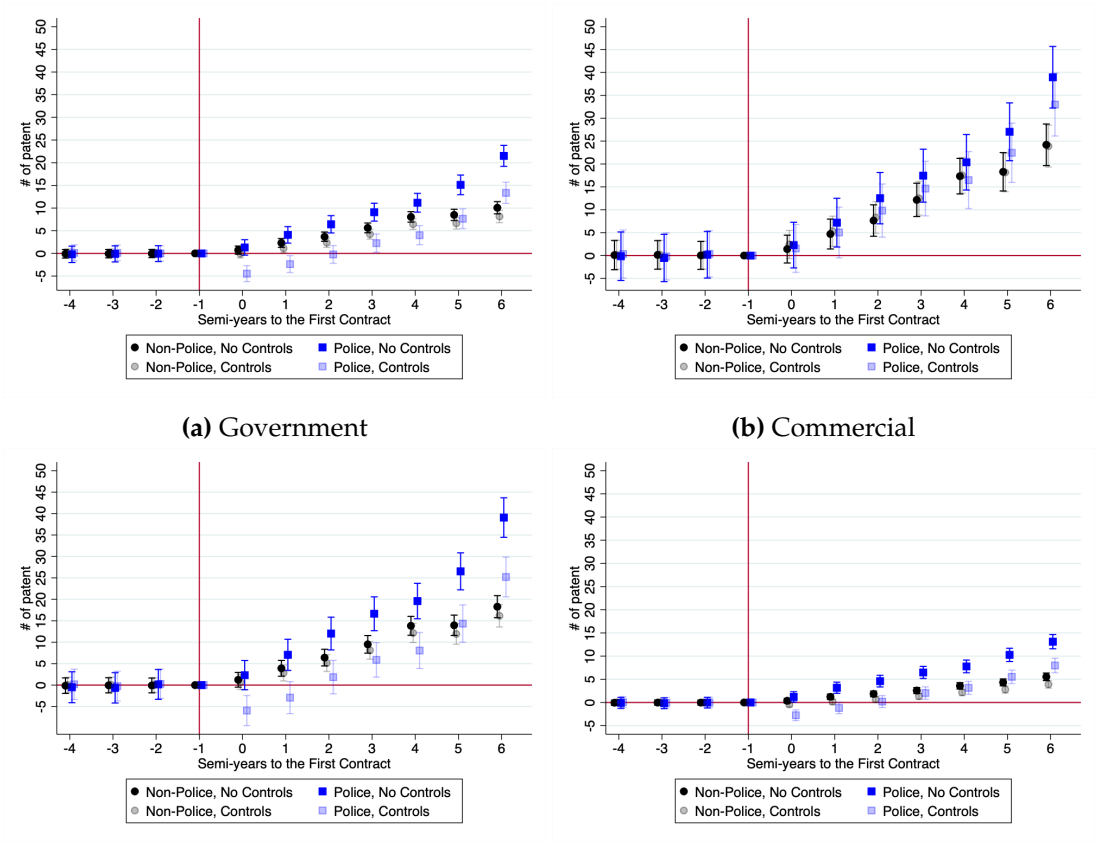
Figure A.12: Classification accuracy of the Recurrent Neural Network with Long Short-Term Memory by year in which IP was released/registered. Top: for patents, Bottom: for Software.



(a) Government **(b)** Commercial

(c) Data-Complementary **(d)** Video-Related AI

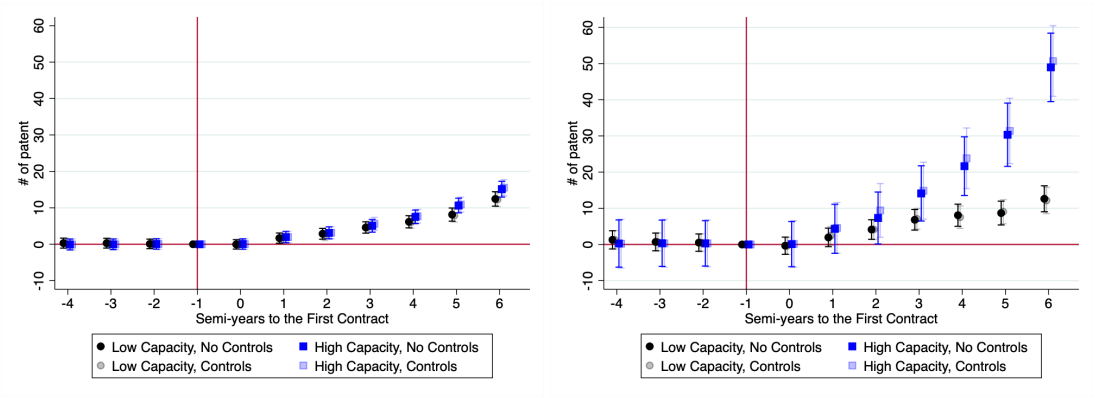
Figure A.13: Software development intended for government (top panel – left), and for commercial uses (top panel – right), relative to the time of receiving initial procurement contract, controlling for firms and time period fixed effects. Data-complementary software development — data storage, data transmission, or data management (bottom panel – left), and video-related AI software (bottom panel – right), relative to the time of receiving initial procurement contract, controlling for firms and time period fixed effects. Blue dots indicate the effects of first public security contract from localities with above median surveillance capacity, and black dots indicate the effects of first public security contract from localities with below median surveillance capacity. Translucent lines/markers additionally interacts outcomes with pre-period variables including area gdp, firm size, and contract size.



(a) Government **(b)** Commercial

(c) Data-Complementary **(d)** Video-Related AI

Figure A.14: Patents development intended for government (top panel – left), and for commercial uses (top panel – right), relative to the time of receiving initial procurement contract, controlling for firms and time period fixed effects. Data-complementary patents development — data storage, data transmission, or data management (bottom panel – left), and video-related AI patents (bottom panel – right), relative to the time of receiving initial procurement contract, controlling for firms and time period fixed effects. Blue dots indicate the effects of first public security contract, and black dots indicate the effects of first non-public security contract. Translucent lines/markers additionally interacts outcomes with pre-period variables including area gdp, firm size, and contract size.



(a) Government **(b)** Commercial

(c) Data-Complementary **(d)** Video-Related AI

Figure A.15: Patents development intended for government (top panel – left), and for commercial uses (top panel – right), relative to the time of receiving initial procurement contract, controlling for firms and time period fixed effects. Data-complementary patents development — data storage, data transmission, or data management (bottom panel – left), and video-related AI patents (bottom panel – right), relative to the time of receiving initial procurement contract, controlling for firms and time period fixed effects. Blue dots indicate the effects of first public security contract from localities with above median surveillance capacity, and black dots indicate the effects of first public security contract from localities with below median surveillance capacity. Translucent lines/markers additionally interacts outcomes with pre-period variables including area gdp, firm size, and contract size.

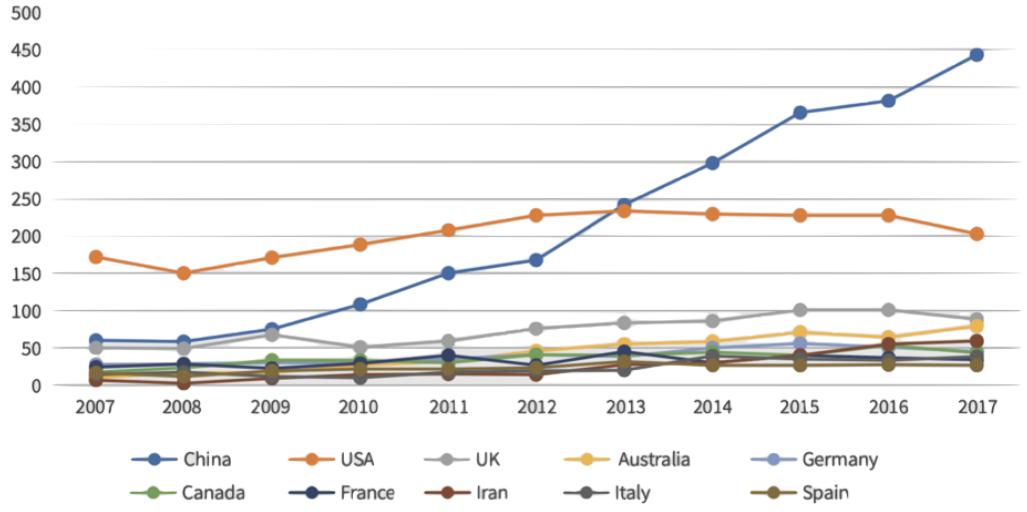


Figure A.16: Highly-cited research in AI over time for selected countries. Source: *China AI Development Report 2018*.

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

Table A.1: Summary Stats - Localities with low versus 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.2: 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.3: Top predicted words from LSTM model — Binary Categorization of Patents

Chinese Term	English Translation	Frequency (%)
(1)	(2)	(3)
Panel A: Customer is Government		
人脸识别	Facial Recognition	.565
交通	Traffic	.535
海信	Hisense	.418
监测	Monitor	.406
腾讯	Tencent	.384
大华	Dahua	.367
安全	Safety	.356
人脸	Face	.340
城市	City	.320
远程	Long Distance/Remote	.217
数字	Numbers	.216
电子	Electronics	.271
安防	Security	.258
影像	Image	.252
DHC	DHC	.251
在线	Online	.240
高清	HD	.222
社区	Community	.218
中心	Center/Hub	.214
车辆	Car	.206
Panel B: Function is AI		
语音	Voice	.610
人脸识别	Facial Recognition	.455
人脸	Face	.361
监测	Monitor	.337
海信	Hisense	.316
交通	Traffic	.281
识别系统	Recognition Systems	.274
车辆	Car	.263
学习	Learning	.252
人工智能	Artificial Intelligence	.245
算法	Algorithm	.235
控制	Control	.221
无人机	Drone	.217
城市	City	.200
DHC	DHC	.191
辅助	Auxilary	.190
三维	3D	.177
医疗	Medical Treatment	.169
动态	Dyamic	.169
交互	Interaction	.167

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			AI-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: Top predicted words from LSTM model — Non-binary Categorization of Patents

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)
飞行器	Aircraft	.427	空调	Air Conditioning	1.804	摄像机	Monitor	1.018
芯片	Chip	.259	空调器	Air Conditioner	1.642	计算机设备	Computer Equipment	.619
数据处理	Data Processing	.231	冰箱	Refrigerator	.146	云台	Tripod Head (Gimbal)	.496
信号	Signal	.229	手机	Mobile Phone	.847	悠悠	Long/Leisurely	.344
目标	Target/Objective	.325	室内	Indoors	.782	测试	Test	.252
视觉	Vision	.203	充电	Charge	.403	终端设备	Terminal Equipment	.251
电池	Battery	.200	制冷	Refrigeration	.355	平台	Platform	.235
通信	Communication	.195	空气	Air	.261	人脸识别	Facial Recognition	.201
搜索	Search	.195	车载	Car	.240	相关	Related	.197
保护	Protection	.186	遥控器	Remote Control	.297	安全	Safety	.189
人脸	Face	.186	游戏	Game	.221	进行	Execute	.184
存储	Storage	.185	电子	Electronics	.204	天线	Antenna	.179
相机	Camera	.184	洗碗机	Dishwasher	.204	交通	Traffic	.178
模型	Model	.180	电脑	Computer	.201	多功能	Multipurpose	.168
播放	Broadcast	.178	交互	Inteaction	.222	安装	Installation	.167
图像处理	Image Processing	.177	压缩机	Compressor	.197	硬盘	Hard Disk	.166
红外	Infrared	.176	洗衣机	Clothes Washer	.194	利用	Usage	.163
同步	Synchronize	.179	壁挂式	Wall-Mounted	.193	一体机	One Machine	.153
三维	3D	.174	使用	Use	.187	录像机	VCR	.151
确定	Determine	.173	自助	Self Help	.182	适用	Applicable	.150

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)
洗碗机	Dishwasher	.409	人脸	Face	.872	空调器	Air Conditioner	2.186
玩具车	Toy Car	.365	目标	Target/Goal	1.207	室内	Indoors	1.086
推送	Push	.345	图像处理	Image Processing	.624	飞行器	Aircraft	.750
POS	POS	.308	跟踪	Track	.463	无人	Umanned	.585
洗衣机	Clothes Washer	.298	人脸识别	Facial Recognition	.672	电池	Battery	.361
冷藏	Refrigeration	.254	特征	Feature	.373	壁挂式	Wall-Mounted	.249
界面	Interface	.240	视觉	Vision	.346	压缩机	Compressor	.246
支付	Pay	.211	深度	Depth	.322	机械	Machine	.235
冷冻	Freezing	.209	神经网络	Neural Networks	.265	室外机	Outdoor Machine	.232
页面	Page	.192	学习	Learning	.253	空气	Air	.220
导航	Navigation	.178	全景	Panoramic	.248	开关	Switch	.195
电源	Power Supply	.168	镜头	Lens	.228	飞行	Aviation	.194
业务	Business	.162	水下	U nderwater	.213	遥控	Remote	.190
柔性	Flexibility	.152	身份	Identity	.212	保护	Protect	.184
手持	Handheld	.149	区域	Region	.201	自助	Self-Help	.180
广告	Advertisement	.144	虚拟现实	Virtual Reality	.259	焊接	Welding	.169
测试方法	test Methods	.140	成像	Imaging	.184	机及	Machine	.164
认证	Certification	.139	触摸	Touch	.172	电梯	Escalator	.157
带有	Involve	.138	技术	Technology	.171	组合	Comprehensive	.141
投影	Projection	.136	交通	Traffic	.171	运行	Run	.134

AI-Common			AI-Complementary		
Chinese	English	Freq. (%)	Chinese	English	Freq. (%)
芯片	Chip	.440	传输	Transmission	.461
推荐	Recommend	.308	存储	Storage	.397
积木	Blocks	.297	输入	Input	.284
文本	Text	.265	触摸屏	Touchscreen	.283
激光雷达	Leidar	.262	同步	Synchronize	.243
公仔	Doll	.244	信号	Signal	.231
图片	Image	.222	快速	High-Speed	.219
指纹识别	Fingerprint Recognition	.304	接入	Access	.193
音频	Audio	.201	红外	Infrared	.189
联网	Network	.187	接口	Interface	.158
验证	Verification	.177	信息处理	Info Processing	.157
虹膜	Touch	.172	LED	LED	.145
监测	Monitor	.169	车灯	Headlight	.145
辅助	Auxillary	.168	旋转	Spin	.144
适应	Adapt	.168	发送	Transmit	.142
预测	Prediction	.160	配置	Configuration	.131
驾驶	Drive	.129	门锁	Door Lock	.129
利用	Usage	.139	数据传输	Data Transmission	.126
光学	Optics	.137	摄像	Video	.126

Table A.6: Public security & local surveillance capacity — software production

	Government	Commercial	AI-comple.	AI-video
	(1)	(2)	(3)	(4)
Panel A: baseline				
4 semi-years before	-0.115	0.008	-0.115	0.021
	(0.196)	(0.131)	(0.159)	(0.036)
6 semi-years after	4.898***	3.692***	4.742***	0.488***
	(0.279)	(0.186)	(0.226)	(0.045)
4 semi-years before × Public security	-0.142	-0.131	-0.083	-0.020
	(0.317)	(0.213)	(0.258)	(0.044)
6 semi-years after × Public security	4.175***	4.238***	4.869***	1.008***
	(0.421)	(0.281)	(0.341)	(0.058)
Panel B: local surveillance capacity				
4 semi-years before	-0.226	-0.035	0.007	0.004
	(0.378)	(0.337)	(0.373)	(0.062)
6 semi-years after	9.354***	6.950***	9.018***	1.392***
	(0.484)	(0.423)	(0.472)	(0.079)
4 semi-years before × High capacity	-0.037	-0.338	-0.386	-0.065
	(0.539)	(0.477)	(0.529)	(0.088)
6 semi-years after × High capacity	0.968	3.989***	2.780***	0.635***
	(0.691)	(0.606)	(0.674)	(0.113)
Semi-year FE	Y	Y	Y	Y
Subsidiary Firm FE	Y	Y	Y	Y

Notes: Panel A replicates baseline specification in Table 3, Columns 1-4; specification includes full set of time indicators and interactions with public security contracts; only selected coefficient estimates are presented. Panel B includes full set of time indicators and interactions with public security contracts with localities that have above median surveillance capacity; only selected coefficient estimates are presented. Standard errors clustered at mother firm level are reported below the coefficients. * significant at 10% ** significant at 5% *** significant at 1%.

Table A.7: Differential Effects of High and Low Capacity Contracts over Non-Police Contracts

	Government	Commercial	AI-complementary	AI-video
	(1)	(2)	(3)	(4)
Panel A: Baseline				
4 Semiyears Before	-0.117	0.007	-0.108	0.003
	(0.193)	(0.137)	(0.156)	(0.026)
6 Semiyears After	4.884***	3.733***	4.712***	0.534***
	(0.275)	(0.194)	(0.222)	(0.036)
4 Semiyears Before × High Capacity	-0.102	-0.236	-0.227	-0.052
	(0.408)	(0.288)	(0.329)	(0.054)
6 Semiyears After × High Capacity	4.452***	6.341***	6.131***	1.226***
	(0.535)	(0.378)	(0.432)	(0.072)
4 Semiyears Before × Low Capacity	-0.162	-0.014	0.085	0.001
	(0.401)	(0.286)	(0.325)	(0.054)
6 Semiyears After × Low Capacity	4.402***	2.816***	3.947***	0.781***
	(0.522)	(0.367)	(0.420)	(0.070)
p, 6 After × High $\neq$ 6 After × Low	0.938	0.000	0.000	0.000
Panel B: Add Pre-Contract Interactions				
4 Semiyears Before	-0.098	0.008	-0.097	0.002
	(0.193)	(0.130)	(0.157)	(0.026)
6 Semiyears After	4.273***	3.853***	4.460***	0.547***
	(0.275)	(0.185)	(0.223)	(0.037)
4 Semiyears Before × High Capacity	-0.145	-0.215	-0.245	-0.045
	(0.407)	(0.273)	(0.329)	(0.054)
6 Semiyears After × High Capacity	3.992***	6.322***	5.768***	1.237***
	(0.534)	(0.359)	(0.433)	(0.072)
4 Semiyears Before × Low Capacity	-0.142	-0.001	0.104	-0.004
	(0.400)	(0.271)	(0.326)	(0.054)
6 Semiyears After × Low Capacity	4.045***	2.746***	3.604***	0.781***
	(0.521)	(0.348)	(0.421)	(0.070)
p, 6 After × High $\neq$ 6 After × Low	0.935	0.000	0.000	0.000

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. In this table, we present results for three mutually exclusive indicators: non-police contracts, low capacity police contracts, and high capacity police contracts (omitted group non-police). Panel B additionally adds controls for first contract area gdp, contract size, and firm size. Standard errors are reported below the coefficients. * significant at 10% ** significant at 5% *** significant at 1%.

Table A.8: Ambiguous Public Security Robustness

	Government	Commercial	AI-complementary	AI-video
	(1)	(2)	(3)	(4)
Panel A: Baseline				
4 Semiyears Before	-0.109	0.008	-0.116	0.003
	(0.196)	(0.138)	(0.157)	(0.027)
6 Semiyears After	4.963***	3.722***	4.792***	0.538***
	(0.279)	(0.197)	(0.224)	(0.039)
4 Semiyears Before × Public Security	-0.154	-0.135	-0.072	-0.021
	(0.317)	(0.225)	(0.255)	(0.044)
6 Semiyears After × Public Security	4.149***	4.378***	4.665***	0.995***
	(0.421)	(0.297)	(0.338)	(0.059)
Panel B: Dropping Ambiguous Companies				
4 Semiyears Before	-0.119	0.015	-0.117	0.001
	(0.200)	(0.131)	(0.161)	(0.027)
6 Semiyears After	5.236***	3.929***	5.031***	0.567***
	(0.286)	(0.188)	(0.231)	(0.039)
4 Semiyears Before × Public Security	-0.176	-0.098	-0.098	-0.028
	(0.330)	(0.218)	(0.267)	(0.045)
6 Semiyears After × Public Security	3.983***	3.945***	4.487***	0.963***
	(0.438)	(0.287)	(0.353)	(0.060)
Panel C: Ambiguous Companies as Non-Public Security				
4 Semiyears Before	-0.119	-0.020	-0.109	0.002
	(0.192)	(0.129)	(0.154)	(0.026)
6 Semiyears After	5.265***	4.256***	5.109***	0.630***
	(0.272)	(0.182)	(0.219)	(0.038)
4 Semiyears Before × Public Security	-0.162	-0.069	-0.099	-0.027
	(0.326)	(0.219)	(0.263)	(0.045)
6 Semiyears After × Public Security	3.939***	3.584***	4.428***	0.899***
	(0.429)	(0.287)	(0.345)	(0.060)

Notes: Panel A replicates baseline specification in Table 3, Columns 1-4; 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 one an ambiguous contract, while Panel C reclassifies these companies as having a first contract that is non-public security. Standard errors clustered at mother firm level are reported below the coefficients. * significant at 10% ** significant at 5% *** significant at 1%.

Table A.9: LSTM Model Robustness

	Government	Commercial	AI-complementary	AI-video
	(1)	(2)	(3)	(4)
Panel A: Baseline (Timestep 20, Embeddings 32, Nodes 32)				
4 Semiyears Before	-0.097	0.015	-0.088	-0.047
	(0.164)	(0.189)	(0.153)	(0.084)
6 Semiyears After	3.292***	4.193***	4.802***	1.800***
	(0.233)	(0.269)	(0.219)	(0.120)
4 Semiyears Before × Public security	-0.075	-0.162	-0.059	-0.059
	(0.266)	(0.309)	(0.249)	(0.137)
6 Semiyears After × Public security	3.149***	5.699***	3.511***	1.915***
	(0.354)	(0.408)	(0.331)	(0.182)
Panel B: Vary Timestep 10, Embeddings 32, Nodes 32				
4 Semiyears Before	-0.092	-0.026	-0.072	-0.031
	(0.192)	(0.189)	(0.134)	(0.107)
6 Semiyears After	4.199***	3.394***	3.405***	2.215***
	(0.275)	(0.269)	(0.191)	(0.152)
4 Semiyears Before × Public security	-0.086	-0.099	-0.123	-0.098
	(0.312)	(0.309)	(0.217)	(0.173)
6 Semiyears After × Public security	3.044***	6.318***	3.427***	1.802***
	(0.416)	(0.408)	(0.289)	(0.231)
Panel C: Timestep 20, Vary Embeddings 16, Nodes 32				
4 Semiyears Before	-0.126	-0.068	-0.107	-0.061
	(0.194)	(0.167)	(0.150)	(0.124)
6 Semiyears After	4.772***	3.536***	4.158***	2.161***
	(0.277)	(0.238)	(0.214)	(0.178)
4 Semiyears Before × Public security	-0.132	-0.117	-0.081	-0.079
	(0.315)	(0.273)	(0.243)	(0.202)
6 Semiyears After × Public security	4.722***	4.955***	3.942***	2.754***
	(0.420)	(0.362)	(0.322)	(0.269)
Panel D: Timestep 20, Embeddings 32, Vary Nodes 16				
4 Semiyears Before	-0.101	-0.048	-0.059	-0.083
	(0.197)	(0.191)	(0.137)	(0.103)
6 Semiyears After	5.120***	3.253***	4.298***	1.724***
	(0.281)	(0.272)	(0.195)	(0.147)
4 Semiyears Before × Public security	-0.148	-0.109	-0.111	-0.007
	(0.320)	(0.312)	(0.222)	(0.168)
6 Semiyears After × Public security	3.494***	5.693***	3.838***	1.503***
	(0.425)	(0.414)	(0.295)	(0.224)

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 3, Columns 1-4 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 exhibits results for the same model in Panel A trained with a timestep of 10 instead; Panel C exhibits results for the same model in Panel A trained with an embedding size of 16 instead; Panel D exhibits results for the same model in Panel A trained with 32 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.10: Robustness — time frame

	Government	Commercial	AI-comple.	AI-video
	(1)	(2)	(3)	(4)
Panel A: baseline				
4 semi-years before	-0.115	0.008	-0.115	0.021
	(0.196)	(0.131)	(0.159)	(0.036)
6 semi-years after	4.898***	3.692***	4.742***	0.488***
	(0.279)	(0.186)	(0.226)	(0.045)
4 semi-years before × Public security	-0.142	-0.131	-0.083	-0.020
	(0.317)	(0.213)	(0.258)	(0.044)
6 semi-years after × Public security	4.175***	4.238***	4.869***	1.008***
	(0.421)	(0.281)	(0.341)	(0.058)
Panel B: full balanced panel				
4 semi-years before	-0.090	-0.089	-0.066	-0.016
	(0.189)	(0.209)	(0.190)	(0.041)
6 semi-years after	5.842***	4.009***	5.119***	0.549***
	(0.202)	(0.222)	(0.202)	(0.043)
4 semi-years before × Public security	-0.109	-0.066	-0.177	-0.004
	(0.278)	(0.308)	(0.278)	(0.060)
6 semi-years after × Public security	4.436***	5.504***	5.890***	1.229***
	(0.296)	(0.325)	(0.295)	(0.064)
Panel C: extended time frame				
5 semi-years before	-0.030	0.065	0.007	0.016
	(0.239)	(0.097)	(0.147)	(0.023)
8 semi-years after	9.198***	6.831***	7.699***	0.662***
	(0.398)	(0.161)	(0.244)	(0.039)
5 semi-years before × Public security	0.145	0.030	0.106	0.039
	(0.119)	(0.160)	(0.150)	(0.049)
8 semi-years after × Public Security	3.410***	3.716***	4.545***	0.726***
	(0.241)	(0.321)	(0.299)	(0.098)
Panel D: quarter as unit of analysis				
12 quarters before × Public security	-0.083	-0.243	-0.209	-0.013
	(0.355)	(0.273)	(0.313)	(0.055)
18 quarters after × Public security	2.958***	5.998***	6.216***	1.168***
	(0.575)	(0.442)	(0.503)	(0.090)
Semi-year FE	Y	Y	Y	Y
Subsidiary Firm FE	Y	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 3, Columns 1-4; Panel B restrict 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 quarter as unit of analysis, instead of semi-year. * significant at 10% ** significant at 5% *** significant at 1%.

Table A.11: Robustness — mother firms versus subsidiaries

	Government	Commercial	AI-comple.	AI-video
	(1)	(2)	(3)	(4)
Panel A: baseline				
4 semi-years before	-0.115	0.008	-0.115	0.021
	(0.196)	(0.131)	(0.159)	(0.036)
6 semi-years after	4.898***	3.692***	4.742***	0.488***
	(0.279)	(0.186)	(0.226)	(0.045)
4 semi-years before × Public security	-0.142	-0.131	-0.083	-0.020
	(0.317)	(0.213)	(0.258)	(0.044)
6 semi-years after × Public security	4.175***	4.238***	4.869***	1.008***
	(0.421)	(0.281)	(0.341)	(0.058)
Panel B: control for mother firms' fixed effects				
4 semi-years before	-0.024	0.008	-0.023	0.007
	(0.219)	(0.149)	(0.177)	(0.029)
6 semi-years after	5.435***	3.951***	5.095***	0.552***
	(0.311)	(0.210)	(0.250)	(0.041)
4 semi-years before × Public security	-0.168	-0.120	-0.108	-0.023
	(0.356)	(0.241)	(0.287)	(0.047)
6 semi-years after × Public security	3.773***	4.530***	4.599***	1.027***
	(0.472)	(0.318)	(0.380)	(0.062)
Panel C: mother firm as unit of analysis				
4 semi-years before	-0.049	0.079	-0.022	0.021
	(0.233)	(0.158)	(0.183)	(0.036)
6 semi-years after	3.668***	3.378***	3.896***	0.488***
	(0.285)	(0.188)	(0.216)	(0.045)
4 semi-years before × Public security	-0.180	-0.248	-0.108	-0.042
	(0.361)	(0.244)	(0.281)	(0.056)
6 semi-years after × Public security	1.728***	2.202***	1.750***	0.516***
	(0.418)	(0.280)	(0.320)	(0.067)
Semi-year FE	Y	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 3, Columns 1-4; Panel B include mother firms' fixed effects (instead of subsidiaries' fixed effects); Panel C uses mother firms as unit of analysis, and the specification include mother firms' fixed effects. * significant at 10% ** significant at 5% *** significant at 1%.

Table A.12: Video-related contracts and subsequent software production

Produced video software for government		Prob. of producing commercial software during semi-year after 1st contract				
Before 1st contract	During 1st semi-year after 1st contract	All AI	Video AI	Non-video AI	Diff. b/w video and non-video	Diff. on produced video for govt.
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A						
_	No	0.72	0.30	0.71	-0.42	-0.06
		(0.45)	(0.46)	(0.45)	(0.51)	(0.06)
_	Yes	0.98	0.60	0.96	-0.36	
		(0.14)	(0.49)	(0.20)	(0.52)	
Panel B						
No	No	0.77	0.41	0.76	-0.34	-0.03
		(0.42)	(0.49)	(0.43)	(0.50)	(0.07)
No	Yes	0.97	0.64	0.96	-0.32	
		(0.16)	(0.48)	(0.20)	(0.50)	
Yes	No	0.69	0.20	0.68	-0.48	0.02
		(0.46)	(0.40)	(0.47)	(0.52)	(0.12)
Yes	Yes	1.00	0.45	0.95	-0.5	
		(0.00)	(0.51)	(0.21)	(0.60)	

Notes: For Columns 3 to 6, standard deviations are reported below the means. For Column 7, standard errors are reported below the means. * significant at 10% ** significant at 5% *** significant at 1%.

Table A.13: Ruling out alternative hypotheses: controls interacting with time trends

	Government	Commercial	AI-complementary	AI-video
	(1)	(2)	(3)	(4)
Panel A: Baseline				
6 Semi-years After	4.897***	3.719***	4.738***	0.536***
	(0.278)	(0.187)	(0.224)	(0.038)
Public security	0.184	0.012	0.222	0.026
	(0.318)	(0.217)	(0.257)	(0.044)
6 Semi-years After × Public security	4.208***	4.274***	4.702***	0.983***
	(0.420)	(0.282)	(0.337)	(0.058)
R^2	0.448	0.418	0.460	0.377
Panel B: Amount of Pre-Contract Software				
6 Semi-years After	1.510***	2.045***	2.473***	0.257***
	(0.182)	(0.288)	(0.233)	(0.039)
Public security	0.286	0.353	0.431	0.060
	(0.211)	(0.330)	(0.267)	(0.045)
6 Semi-years After × Public security	2.931***	1.832***	3.026***	0.809***
	(0.275)	(0.436)	(0.352)	(0.060)
R^2	0.435	0.392	0.406	0.322
Panel C: Firm Pre-Contract Size				
6 Semi-years After	4.077***	4.931***	4.906***	0.520***
	(0.201)	(0.311)	(0.249)	(0.039)
Public security	0.192	0.241	0.308	0.048
	(0.233)	(0.356)	(0.286)	(0.045)
6 Semi-years After × Public security	4.409***	3.636***	4.599***	0.902***
	(0.303)	(0.471)	(0.377)	(0.060)
R^2	0.306	0.290	0.307	0.273
Panel D: Contract Size				
6 Semi-years After	3.774***	5.178***	4.789***	0.555***
	(0.210)	(0.310)	(0.249)	(0.041)
Public security	0.172	0.205	0.280	0.045
	(0.245)	(0.356)	(0.287)	(0.047)
6 Semi-years After × Public security	4.372***	3.653***	4.678***	1.016***
	(0.318)	(0.470)	(0.378)	(0.062)
R^2	0.304	0.293	0.316	0.269
Panel E: First Contract Location GDP				
6 Semi-years After	4.289***	4.647***	4.932***	0.615***
	(0.200)	(0.310)	(0.249)	(0.042)
Public security	0.151	0.218	0.264	0.039
	(0.232)	(0.355)	(0.286)	(0.048)
6 Semi-years After × Public security	4.428***	3.640***	4.602***	1.066***
	(0.302)	(0.469)	(0.377)	(0.063)
R^2	0.308	0.293	0.308	0.271
Panel F: All Pre-controls Interacted				
6 Semi-years After	2.355***	1.410***	2.570***	0.299***
	(0.181)	(0.288)	(0.232)	(0.040)
Public security	0.247	0.374	0.399	0.055
	(0.209)	(0.328)	(0.264)	(0.046)
6 Semi-years After × Public security	3.119***	1.773***	2.950***	0.843***
	(0.272)	(0.434)	(0.348)	(0.060)
R^2	0.445	0.396	0.415	0.324

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 3, Columns 1-4. Panels B through F interacts controls with time dummies, where Panel B interacts the amount of software produced by the firm prior to the first contract, Panel C interacts the monetary size of the firm, Panel D interacts the size of the contract, Panel E interacts the GDP of the location of the first contract, and Panel F interacts with all the above controls. 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.14: Increases in software and patents development after 2nd contracts

	Government	Commercial	AI-comple.	AI-video
	(1)	(2)	(3)	(4)
Panel A: software production				
4 semi-years before	0.047	-0.000	0.090	0.001
	(0.126)	(0.208)	(0.184)	(0.040)
6 semi-years after	5.756***	5.218***	6.049***	0.594***
	(0.199)	(0.331)	(0.289)	(0.064)
4 semi-years before × Public security	-0.231	0.001	-0.168	-0.029
	(0.204)	(0.337)	(0.296)	(0.065)
6 semi-years after × Public security	6.174***	8.349***	8.437***	1.726***
	(0.306)	(0.505)	(0.443)	(0.098)
Panel B: patents development				
4 semi-years before	-0.382	-0.570	-0.709	-0.287
	(0.741)	(2.571)	(1.598)	(0.484)
6 semi-years after	17.606***	19.262***	23.408***	11.102***
	(1.173)	(4.089)	(2.516)	(0.775)
4 semi-years before × Public security	-0.136	0.575	-0.091	-0.058
	(1.204)	(4.164)	(2.574)	(0.788)
6 semi-years after × Public security	32.480***	103.421***	72.858***	15.602***
	(1.802)	(6.243)	(3.860)	(1.179)
Semi-year FE	Y	Y	Y	Y
Subsidiary Firm FE	Y	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. * significant at 10% ** significant at 5% *** significant at 1%.

Table A.15: Capital injection, scale effects, and learning-by-doing

	Government	Commercial	AI-comple.	AI-video
	(1)	(2)	(3)	(4)
Panel A: baseline				
4 semi-years before	-0.115	0.008	-0.115	0.021
	(0.196)	(0.131)	(0.159)	(0.036)
6 semi-years after	4.898***	3.692***	4.742***	0.488***
	(0.279)	(0.186)	(0.226)	(0.045)
4 semi-years before × Public security	-0.142	-0.131	-0.083	-0.020
	(0.317)	(0.213)	(0.258)	(0.044)
6 semi-years after × Public security	4.175***	4.238***	4.869***	1.008***
	(0.421)	(0.281)	(0.341)	(0.058)
Panel B: subsample of firms with above median capitalization prior to 1st contract				
4 semi-years before	-0.064	-0.035	-0.054	-0.003
	(0.269)	(0.174)	(0.217)	(0.034)
6 semi-years after	6.503***	4.319***	5.667***	0.601***
	(0.362)	(0.234)	(0.293)	(0.045)
4 semi-years before × Public security	-0.164	-0.152	-0.161	-0.000
	(0.423)	(0.275)	(0.340)	(0.053)
6 semi-years after × Public security	3.771***	4.894***	5.144***	1.074***
	(0.538)	(0.348)	(0.433)	(0.068)
Panel C: control for lagged software production in corresponding categories				
4 semi-years before	-0.005	-0.005	-0.005	-0.000
	(0.024)	(0.020)	(0.024)	(0.006)
6 semi-years after	0.161***	0.237***	0.211***	0.028***
	(0.034)	(0.028)	(0.033)	(0.008)
4 semi-years before × Public security	-0.004	-0.006	0.001	-0.002
	(0.039)	(0.032)	(0.038)	(0.010)
6 semi-years after × Public security	0.469***	0.293***	0.384***	0.136***
	(0.051)	(0.042)	(0.051)	(0.013)
Panel D: control for lagged software production in opposite categories				
4 semi-years before	0.174	0.066	0.107	0.016
	(0.130)	(0.135)	(0.109)	(0.027)
6 semi-years after	-0.075	2.016***	0.949***	0.281***
	(0.184)	(0.191)	(0.155)	(0.039)
4 semi-years before × Public security	-0.255	-0.145	-0.234	-0.031
	(0.211)	(0.219)	(0.177)	(0.044)
6 semi-years after × Public security	2.748***	4.140***	3.368***	0.914***
	(0.280)	(0.289)	(0.234)	(0.059)
Semi-year FE	Y	Y	Y	Y
Subsidiary Firm FE	Y	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. Opposite categories (Panel D) means the pairs 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.16: Robustness — firms' capitalization

	Government	Commercial	AI-comple.	AI-video
	(1)	(2)	(3)	(4)
Panel A: baseline				
4 semi-years before	-0.115	0.008	-0.115	0.021
	(0.196)	(0.131)	(0.159)	(0.036)
6 semi-years after	4.898***	3.692***	4.742***	0.488***
	(0.279)	(0.186)	(0.226)	(0.045)
4 semi-years before × Public security	-0.142	-0.131	-0.083	-0.020
	(0.317)	(0.213)	(0.258)	(0.044)
6 semi-years after × Public security	4.175***	4.238***	4.869***	1.008***
	(0.421)	(0.281)	(0.341)	(0.058)
Panel B: trim firms with smallest capitalization				
4 semi-years before	0.015	0.066	0.010	0.001
	(0.263)	(0.168)	(0.207)	(0.030)
6 semi-years after	6.885***	5.069***	6.429***	0.567***
	(0.395)	(0.250)	(0.310)	(0.046)
4 semi-years before × Public security	-0.365	-0.289	-0.321	-0.010
	(0.559)	(0.358)	(0.439)	(0.066)
6 semi-years after × Public security	6.698***	7.916***	8.056***	1.541***
	(0.681)	(0.433)	(0.533)	(0.080)
Panel C: weighted by firm capitalization				
4 semi-years before	-0.150	0.902***	0.381***	0.094***
	(0.127)	(0.092)	(0.140)	(0.015)
6 semi-years after	23.548***	12.738***	19.114***	0.888***
	(0.148)	(0.108)	(0.164)	(0.017)
4 semi-years before × Public security	0.134	-0.853***	-0.400**	-0.014
	(0.180)	(0.131)	(0.198)	(0.021)
6 semi-years after × Public security	-2.913***	10.848***	9.362***	1.391***
	(0.199)	(0.145)	(0.221)	(0.023)
Semi-year FE	Y	Y	Y	Y
Subsidiary Firm FE	Y	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 3, Columns 1-4; Panel B trims the sample by the extreme tail of firms with smallest capitalization; Panel C weighs the baseline specification with firms capitalization prior to 1st contracts. * significant at 10% ** significant at 5% *** significant at 1%.

Appendix A Proofs

Appendix A.1 Proof of Proposition 1

We have that the profit maximization problem is

$$\max_{d_p, \overrightarrow{x}_p, \overrightarrow{x}_g} \left((d_g)^{\beta} X(\overrightarrow{x}_g)^{1-\beta} \right)^{\frac{\chi-1}{\chi}} + \left((\alpha d_g^{\frac{\gamma-1}{\gamma}} + (1-\alpha) d_p^{\frac{\gamma-1}{\gamma}})^{\frac{\gamma}{\gamma-1}\beta} X(\overrightarrow{x}_p)^{1-\beta} \right)^{\frac{\chi-1}{\chi}} - w_p d_p - \sum w_i (x_{gi} + x_{pi})$$

The FOC are

$$\begin{aligned}
\frac{\chi-1}{\chi} (q_p)^{\frac{\chi-1}{\chi}} \beta \frac{(1-\alpha) d_p^{-\frac{1}{\gamma}}}{\alpha d_g^{\frac{\gamma-1}{\gamma}} + (1-\alpha) d_p^{\frac{\gamma-1}{\gamma}}} &= w_d \\
\frac{\chi-1}{\chi} (q_p)^{\frac{\chi-1}{\chi}} (1-\beta) \frac{X_{pi}}{X} &= w_i \\
\frac{\chi-1}{\chi} (q_g)^{\frac{\chi-1}{\chi}} (1-\beta) \frac{X_{gi}}{X} &= w_i
\end{aligned}$$

Also,

$$\sigma = \frac{\alpha d_g^{\frac{\gamma-1}{\gamma}}}{\alpha d_g^{\frac{\gamma-1}{\gamma}} + (1-\alpha) d_p^{\frac{\gamma-1}{\gamma}}}$$

Totally differentiating the private software production technology and the FOCs

$$\begin{aligned}
\Delta log(q_p) &= \beta\sigma \Delta log(d_g) + \beta(1-\sigma)\Delta log(d_p) + (1-\beta) \sum \frac{X_i x_{pi}}{X} \Delta log(x_{pi}) \\
\Delta log(d_p) &= \frac{\frac{\chi-1}{\chi} \Delta log\left(q_p\right) - \sigma \Delta log(d_g)}{(1 + \frac{1}{\gamma} - \sigma)} \\
\Delta log(x_{pi}) &= \frac{\chi-1}{\chi} \Delta log\left(q_p\right) \sum \frac{X_i x_{pi}}{X} + \sum \Delta \left(\frac{X_i x_i}{X} \right)
\end{aligned}$$

Under the assumption that X is constant returns to scale, we have that $\sum \frac{X_i x_{pi}}{X} = 1$ and thus $\sum \Delta \left(\frac{X_i x_i}{X} \right) = 0$. Using above and combining the expressions we obtain

$$\Delta log(q_p) = \frac{\frac{1}{\gamma}\beta\sigma}{1 + \frac{1}{\gamma} - \sigma} \Delta log(d_g) + \frac{(1-\sigma) + (1-\beta)\frac{1}{\gamma}}{1 + \frac{1}{\gamma} - \sigma} \frac{\chi-1}{\chi} \Delta log\left(q_p\right)$$

Therefore,

$$\Delta log(q_p) = \frac{\chi\beta\sigma}{1 + (\chi-1)\beta + \gamma(1-\sigma)} \Delta log(d_g)$$

which is the expression in the proposition.

The proof for the expression for $\Delta log(q_g)$ is analogous except that $\sigma = 1$. Finally, dividing both expressions and re-organizing terms we obtain the expression for σ in the proposition.

Appendix A.2 Existence and uniqueness of a BGP equilibrium with entry of both types of firms

Proposition 2 (Existence and Uniqueness) *Let $p_z(p_p)$ be the implicit function defined by the pricing equation* (4) *and $\underline{p}_p$ the smallest p_p such that $p_z(\underline{p}_p) \geq 0$. Let $p_d(p_p)$ be the implicit function defined by*

$$\Pi_p(0, p_p, p_d) = \mu_z \Pi(p_z(p_p)). \tag{30}$$

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

$$\kappa_g \frac{\Pi(p_g, \bar{d}_g)}{p_g} \frac{\chi}{1 + \beta(\chi - 1)} = \bar{d}_g. \tag{31}$$

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

$$\frac{q_p(0, p_p, p_d(p_p))^{1-\frac{1}{\chi}}}{d_p(0, p_p, p_d(p_p))} < \frac{Y_p}{D_p} = \frac{\left(\frac{p_p}{1-a}\right)^{-\epsilon}}{\left((1-\psi)(p_d(p_p)(1-\tau))\right)^{\frac{1-\psi}{\psi}} \kappa_p} < \frac{q_p(\bar{d}_g, p_p, p_d(p_p))^{1-\frac{1}{\chi}}}{d_p(\bar{d}_g, p_p, p_d(p_p))} \tag{32}$$

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

$$\gamma \geq 1 + \beta(\chi - 1) \tag{33}$$

$$\Pi_g(p_g(\bar{d}_g), \bar{d}_g) + \Pi_p(\bar{d}_g, \underline{p}_p, p_d(\underline{p}_p)) - \Pi_p(0, \underline{p}_p, p_d(\underline{p}_p)) < \frac{F}{\lambda} \tag{34}$$

We no proceed to prove this proposition. Equations (30) and (31) follow directly from the free-entry conditions of private innovators and market clearing in the government data and goods markets. It is easy to show that $p_z(p_p)$ exist and has a negative derivative, and that $p_d(p_p)$ 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 $\tilde{N}_p / N_z, N_g / N_z$ as a function of p_p alone.

$$\begin{bmatrix} \frac{\tilde{N}_p}{N_z} \\ \frac{N_g}{N_z} \end{bmatrix} = \begin{bmatrix} q_p(0, p_p, p_d(p_p))^{1-\frac{1}{\chi}} & q_p(\bar{d}_g, p_p, p_d(p_p))^{1-\frac{1}{\chi}} \\ d_p(0, p_p, p_d(p_p)) & d_p(\bar{d}_g, p_p, p_d(p_p)) \end{bmatrix}^{-1} \begin{bmatrix} \frac{Y_p}{N_z} \\ \frac{D_p}{N_z} \end{bmatrix}$$

$$\begin{bmatrix} \frac{Y_p}{N_z} \\ \frac{D_p}{N_z} \end{bmatrix} = \begin{bmatrix} \left(\frac{p_p}{1-a}\right)^{-\epsilon} \\ \left((1-\psi)(p_d(p_p)(1-\tau))\right)^{\frac{1-\psi}{\psi}} \kappa_p \end{bmatrix} \left(\frac{p_z(p_p)}{a}\right)^{\epsilon} q_z(p_z(p_p))^{1-\frac{1}{\chi}}$$

When the determinant of the square matrix is negative, then $\tilde{N}_p/N_z > 0$ and $N_g/N_z > 0$ if and only if the inequalities in (32) hold. We now show that the determinant is indeed negative. This requires showing that

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

which is also necessary for (32) to hold. The optimality condition for private data demand is,

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

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

$$\frac{q_p(\bar{d}_g, p_p, p_d(p_p))^{1-\frac{1}{\chi}}}{d_p(\bar{d}_g, p_p, p_d(p_p))} = \frac{\chi}{\chi-1}\frac{p_d(p_p)}{\beta p_p}\left(\frac{\alpha}{(1-\alpha)}\left(\frac{\bar{d}_g}{d_p(\bar{d}_g, p_p, p_d(p_p))}\right)^{\frac{\gamma-1}{\gamma}}+1\right) \tag{36}$$

$$= \frac{q_p(0, p_p, p_d(p_p))^{1-\frac{1}{\chi}}}{d_p(0, p_p, p_d(p_p))}\left(\frac{\alpha}{(1-\alpha)}\left(\frac{\bar{d}_g}{d_p(\bar{d}_g, p_p, p_d(p_p))}\right)^{\frac{\gamma-1}{\gamma}}+1\right) \tag{37}$$

$$> \frac{q_p(0, p_p, p_d(p_p))^{1-\frac{1}{\chi}}}{d_p(0, p_p, p_d(p_p))} \tag{38}$$

To conclude the proof, we need to show conditions under which p_p 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_p

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

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_p. Totally differentiating

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

The second equality follows from the implicit function $p_d(p_p)$, the second equation from the envelope theorem, and the third equation from equation (36). The last inequality follows from (35) and the fact that $\frac{\partial \Pi_z(p_z)}{\partial p_z}\frac{\partial p_z}{\partial p_p} < 0$. It is easy to see that when $\gamma \geq 1 + \beta(\chi - 1)$, the $d_p(\bar{d}_g, p_p, p_d)$ is weakly decreasing in $\bar{d}_g$. As such, $\frac{d_p(\bar{d}_g, p_p, p_d)}{d_p(0, p_p, p_d)} \leq 1$ and the inequality holds.

Finally, since when $\gamma > 1 + \beta(\chi - 1)$ the LHS is increasing in p_p, Bolzano's theorem implies that a necessary and sufficient condition for p_p to exist and be unique is that the LHS evaluated at the lowest p_p is lower than F/λ. A sufficient condition for this to occur is equation (33), which completes the proof.

Appendix A.3 Proof of Theorem 1

We first show the comparative statics of η and n_p with respect to changes in $\bar{d}_g$. We then show that $p_g G/Y$ increases whenever $\bar{d}_g$ increases.

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

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

We have shown in the proof of Theorem 2 that the denominator in $\frac{\partial p_p}{\partial \bar{d}_g}$ 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

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

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

$$\begin{aligned}
\frac{Y_p}{Y_z} &= \frac{\tilde{N}_p}{N_z} \frac{1}{q_z(p_z)^{\frac{\chi-1}{\chi}}} q(0, p_p, p_d)^{\frac{\chi-1}{\chi}} + \frac{N_g}{N_z} \frac{1}{q_z(p_z)^{\frac{\chi-1}{\chi}}} q(\bar{d}_g, p_p, p_d)^{\frac{\chi-1}{\chi}} \\
&= \frac{N_p}{N_z} \frac{q(0, p_p, p_d)^{\frac{\chi-1}{\chi}} + q(\bar{d}_g, p_p, p_d)^{\frac{\chi-1}{\chi}}}{q_z(p_z)^{\frac{\chi-1}{\chi}}} \\
&= \frac{N_p}{N_z} \frac{\Pi_p(0, p_p, p_d)/p_p + \Pi_p(\bar{d}_g, p_p, p_d)/p_p \frac{1+\beta(\chi-1)}{1+\beta(\chi-1)\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(\bar{d}_p)^{\frac{\gamma-1}{\gamma}}}}}{\Pi_z(p_z)/p_z} \\
&= \frac{p_z}{p_p} \frac{N_p}{N_z} \left(\mu_z + \frac{\Pi_p(\bar{d}_g, p_p, p_d)}{\Pi_z(p_z)} \frac{1+\beta(\chi-1)}{1+\beta(\chi-1)\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(\bar{d}_p)^{\frac{\gamma-1}{\gamma}}}} \right) \\
&= \frac{p_z}{p_p} \frac{N_p}{N_z} \left(\mu_z + 2\frac{\Pi_p(\bar{d}_g, p_p, p_d)}{\Pi_p(\bar{d}_g, p_p, p_d) + \Pi_g(\bar{d}_g, p_p, p_d) - \frac{F}{\lambda}} \frac{1+\beta(\chi-1)}{1+\beta(\chi-1)\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(\bar{d}_p)^{\frac{\gamma-1}{\gamma}}}} \right)
\end{aligned}$$

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 $\frac{\Pi_p(\bar{d}_g, p_p, p_d)}{\Pi_p(\bar{d}_g, p_p, p_d) + \Pi_g(\bar{d}_g, p_p, p_d) - \frac{F}{\lambda}}$ also decreases. We have that derivative is

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

We know that $\frac{\partial}{\partial \bar{d}_g}\Pi_p(0, p_p(\bar{d}_g), p_d(\bar{d}_g)) > 0$, since $\Pi_z(p_z)$ increases with $\bar{d}_g$ and, in a BGP, this is equal to $\frac{1}{\mu_z}\Pi_p(0, p_p, p_d)$. Then it has to be that $\frac{\partial}{\partial \bar{d}_g}\Pi_p(\bar{d}_g, p_p(\bar{d}_g), p_d(\bar{d}_g)) > 0$, 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 (33) holds.

Finally, from the goods demand equations

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

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

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

Part 3. Government Spending We have that

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

where the second equality follows from the solution to N_g/N_z in Theorem 2 and the last equality uses the expressions in (36).

The first term inside the parenthesis is increasing in $\bar{d}_g$ whenever $\epsilon \geq 1$, since we have shown that p_p 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 $\frac{\bar{d}_g}{d_p(\bar{d}_g,p_p,p_d(p_p))}$ increases with $\bar{d}_g$. This is true when $d_p(\bar{d}_g, p_p, p_d(p_p))$ 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_p. 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. Not also that not only $p_g G/Y$ increases but also G/Y. This then completes the proof.

Appendix A.4 Proof of Theorem 2

From the converse of Theorem 1, we know that when $\bar{d}_g$ decreases then $\eta, n_p, p_g G/Y$ decrease too. Then, to prove the first part of Theorem 2, we simply have to show that in such case D_g/Y decrease as well. This follows directly from the fact that $D_g/Y = \kappa_g G/Y$ and we have shown that G/Y increases.

Appendix A.5 Proof of Theorem 3

The proof for the first part of theorem regarding the effects of $\bar{d}_g$ is identical to the proof in Theorem 1.

As for the the effects of a subsidy to private data, we begin by noting that the BGP equilibrium in Theorem 2 gives a unique set of prices p_d, p_p, p_z, r that are independent of τ. This immediately implies that η is independent of τ. Moreover, from the market clearing equations in private goods and data markets, we see that $\tilde{N}_p/N_z$ and N_g/N_z depend on τ for the given set of prices. Yet, n_p only depends on the prices and not on τ, and is thus independent of it. Finally, since $D_p/Y = \kappa_p((1-\psi)p_d(1-tau)^{(1-\psi)/\psi}$ and p_d is independent of τ, we see that D_p/Y decreases with τ.

Appendix A.6 Equilibrium conditions for quantitative analysis

The maximization problem of software producers is

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

FOC are

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

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

This implies

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

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

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

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

which then gives

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

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

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

So, normalize $\phi = (1-\beta)$. We then get that.

$$\Pi_g(\bar{d}_g, p_g) = (p_g)^{\frac{\chi}{1+\beta(\chi-1)}} (\bar{d}_g)^{\frac{\beta(\chi-1)}{1+\beta(\chi-1)}} \frac{1+\beta(\chi-1)}{\chi-1}$$
$$Y_g = N_g \frac{\Pi_g(\bar{d}_g, p_g)}{p_g}\frac{\chi}{1+\beta(\chi-1)}$$
$$D_g = N_g \bar{d}_g$$
$$\Pi_p(\bar{d}_g, p_p, p_d) = (p_p)^{\frac{\chi}{1+\beta(\chi-1)}} \left(\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(\bar{d}_p)^{\frac{\gamma-1}{\gamma}}\right)^{\frac{\gamma}{\gamma-1}\frac{\beta(\chi-1)}{1+\beta(\chi-1)}} \frac{1}{\chi-1} \times \cdots$$
$$\left(1+\beta(\chi-1)\frac{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}}}{\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(\bar{d}_p)^{\frac{\gamma-1}{\gamma}}}\right)$$
$$(\bar{d}_p)^{\frac{1}{\gamma}} = \frac{(1-\alpha)}{p_d}(p_p)^{\frac{\chi}{1+\beta(\chi-1)}} \left(\alpha(\bar{d}_g)^{\frac{\gamma-1}{\gamma}} + (1-\alpha)(\bar{d}_p)^{\frac{\gamma-1}{\gamma}}\right)^{\frac{\gamma}{\gamma-1}\frac{\beta(\chi-1)}{1+\beta(\chi-1)}-1} \beta$$
$$\Pi_p(0, p_p, p_d) = (p_p)^{\frac{\chi}{1+\beta(\chi-1)}} \left((1-\alpha)^{\frac{\gamma}{\gamma-1}}\underline{d}_p\right)^{\frac{\beta(\chi-1)}{1+\beta(\chi-1)}} \frac{1}{\chi-1}$$

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

Furthermore, from the profit maximization of the final goods seller we get

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

And the market clearing conditions for are

$$
\begin{aligned}
G &= Y_g \\
D_g &= \kappa_g G \\
D_p &= \left((1-\psi)(1-\tau)p_d\right)^{\frac{1-\psi}{\psi}} Y
\end{aligned}
$$

And the free entry conditions are

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

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

Appendix B Intuitions of theorems

Appendix B.1 Theorem 1: rate and direction of innovation in an AI-tocracy

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 an increase in $p_g G/Y$, which endogenously brings about these changes in equilibrium). 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 non-rival.

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_p, 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_p 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_p 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_p, 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_p and to show that $\bar{d}_g, p_g$ indeed increase with $p_g G/Y$. From the definition of n_p together with equations (20) and (21), we obtain:

$$n_p = \underbrace{\left(\frac{1-a}{a}\frac{p_z}{p_p}\right)^{\epsilon}}_{=\frac{Y_p}{Y_z}} \frac{q_z (p_z)^{\frac{\chi-1}{\chi}}}{q(0,p_p,p_d)^{\frac{\chi-1}{\chi}} + q(\bar{d}_g,p_p,p_d)^{\frac{\chi-1}{\chi}}} \tag{39}$$

Thus, there are two countervailing effects on direction of private innovation from the increase in $\bar{d}_g, p_g$. First, the increase in p_z and decrease in p_p result in an increase in the relative demand for private software $\frac{Y_p}{Y_z}$. This demand effect biases the direction of innovation more towards private software (increases n_p). Second, the combined increase in $\bar{d}_g$ and changes in p_p, p_d result in an increase in the relative output of private software per

firm (the second term decreases). This decreases n_p. 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_p increases.

To conclude the intuition for the theorem, consider the market clearing condition for government data (24). 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 $\bar{d}_g$. Finally, the theorem shows that increases in $\bar{d}_g$ are associated with increases in $p_g G/Y$ in a BGP as well.

Appendix B.2 Theorem 3: industrial policy

To understand the neutrality of η and n_p with respect to changes in τ, it helps to go back to the intuition behind Theorem 1 as well. We showed there that prices p_p, p_z, the interest rate r, and, in particular, the private data price *paid by innovators* p_d are fully determined given $\bar{d}_g$ from the free-entry conditions in a BGP. Yet, such free-entry conditions and government data provision $\bar{d}_g$ are independent of τ. Since η and n_p only depend on such prices, r, and $\bar{d}_g$, the neutrality result follows.

Economically, the above discussion implies that private data demand is infinitely elastic in a BGP equilibrium since p_d is not affected by changes in the quantity of private data D_p/Y. However, the lower tax τ does shift the private data supply schedule. Thus, the price received by data producers increases and so does D_p/Y in a BGP equilibrium. But how is such increase in data supply consistent with an equilibrium if neither n_p nor η change? From the market clearing conditions for private data and software, we see that what has to adjust are the number of private software varieties produced by private software only firms $\tilde{N}_p$ relative to those produced by firms also producing for the government N_g. In particular, the changes need to be such that the supply of private software remains unchanged (since the demand has not changed, as prices remain the same pos-tax increase), but the demand of private data increases. Since private data demand (private software production) from private software only firms is higher (lower) than from government software producing firms, then the necessary changes are achieved by an increase in $\tilde{N}_p/N_z$ and a decrease in N_g/N_z that nevertheless result in the the same n_p.

When $G = \bar{d}_g = 0$, our model reduces to a standard two-sector directed-technical change model where the common view of industrial policy holds. The demand for private data (the key input subsidized) is no longer infinitely elastic in a BGP and, thus, prices p_p, p_z, p_d are no longer determined independently of the direction of innovation $\tilde{N}_p/N_z$. The reason is that, as opposed to Theorem 3, the number of varieties N_g/N_z can no longer be adjusted – since it is zero without any government surveillance demand – to accommodate the increase in private data supply brought about by the subsidy.

When $D_g = \kappa_g G + \vartheta Y$, again private demand demand is not infinitely elastic in a BGP. The reason is that the price of government data p_g is no longer determined uniquely by $\bar{d}_g$, as we discussed in Theorem 1. Instead, because the increase in the subsidy to private

data encourages entry of private software firms ($\tilde{N}_p/N_z$ increases), this then increases the volume of private transactions as measured by Y and results in a shift in the supply of government data. Ultimately, $\bar{d}_g$ needs to increase to accommodate the increase in government data supply and/or p_g needs to fall in other to lower the production of surveillance and thus the supply of government data. Either change affects the rate and direction of innovation, but differently so. Thus, the effect on macroeconomic outcomes will ultimately depend on what is it that the state policy decides to keep fixed after the subsidy increase: the total amount of surveillance spending $p_g G/Y$ as in Panel B.2 of Table 4 or the amount of data provide to firms $\bar{d}_g$. Moreover, the consequences of private data subsidies crucially depend on how elastic the supply of private data is (as measured by $1/\psi$). We next explore some of these issues quantitatively.

Appendix C Details on quantitative exercises

To get a sense of the magnitude of autocratic advantage in data-intensive innovation, we calibrate our model. We externally calibrate $\theta = 1, \rho = 0.03, \chi = 3$, 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 ($\frac{\tilde{N}_p}{N_z} = \frac{N_g}{N_z} = 1$) and all sectors have an identical output share ($\frac{p_p Y_p}{p_z Y_z} = \frac{p_g G}{p_z Y_z} = 1$). We assume a growth rate of 2 percent, which matches the average historical growth rate in the US.

The parameters left to set are those associated with data as an input in innovation: the elasticity of private data supply ψ, 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 ψ, β and γ. Our empirical evidence on the responses of government and private software to obtaining government contracts with more data involved only shows that $\beta > 0$ and $\gamma < \infty$. As such, we will externally calibrate them and perform sensitivity analysis. For our benchmark, we set $\beta = 0.4$, $\gamma = 4$, and $\psi = 1$.

However, given β, γ, Proposition 1 pins down the degree of non-rivalry σ from our empirical evidence. From Table 3's responses at a 2 year horizon, the elasticity of private to government software is 0.6, which implies $\sigma = 0.83$. Then, we set α to match σ. In particular, $\alpha = \sigma$ when we normalize the government to private data in software production in the initial BGP to be identical ($\bar{d}_g = d_p(\bar{d}_g, p_p, p_d)$). We will again perform sensitivity analysis for different values of σ consistent with our empirical evidence.