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

Martin Beraja  David Yang  Noam Yuchtman
MIT           Harvard           LSE

Stanford Macro Seminar
September 14th, 2020

Research assistants: Haoran Gao, Shiyun Hu, Andrew Kao, Shuhao Lu, Junxi Liu, Shengqi Ni, Wenwei Peng, Yucheng Quan, Linchuan Xu, Peilin Yang, and Guoli Yin
Motivation: data as input in AI innovation

▶ AI technologies are increasingly widespread
   (Aghion et al. 2017; Acemoglu and Restrepo 2018; Agrawal et al. 2018; Jones and Tonetti 2019)
Motivation: data as input in AI innovation

- AI technologies are increasingly widespread
  (Aghion et al. 2017; Acemoglu and Restrepo 2018; Agrawal et al. 2018; Jones and Tonetti 2019)

- AI innovation is data-intensive
  - In addition to algorithmic advances, data availability has played a significant role in determining the rate of technological progress in AI since the Deep Learning Revolution (Sejnowski, 2018)
  - E.g., facial recognition, speech recognition, chess mastery, and translation — AI advances made with decades-old algorithms applied to recently collected big data (Edge.org; Kai-Fu Lee)
Our argument

Direction of innovation and growth in data-intensive economies may be crucially shaped by the state.
Our argument

Direction of innovation and growth in data-intensive economies may be crucially shaped by the state

- Data differs from traditional inputs in two ways:
  1. **States** are key collectors and repositories of data
     - Tellingly, “state” is at the root of “statistics” (Scott 1998)
  2. Data can be **shared** across multiple uses within the firm
Our argument

Direction of innovation and growth in data-intensive economies may be crucially shaped by the state

- Data differs from traditional inputs in two ways:
  1. **States** are key collectors and repositories of data
     - Tellingly, “state” is at the root of “statistics” (Scott 1998)
  2. Data can be **shared** across multiple uses within the firm

- Thus, **economies of scope** may arise from government data
  - An AI firm gaining access to government data could use it develop software for both **government** and **commercial** uses
1. Are there economies of scope arising from government data?
1. Empirically study a prototypical data-intensive industry with significant state interest: facial recognition AI in China
This paper

1. Empirically study a prototypical data-intensive industry with significant state interest: facial recognition AI in China

2. What are the macro implications of firms’ access to government data?
This paper

1. Empirically study a prototypical data-intensive sector with significant state interest: **facial recognition AI in China**

2. **Directed technical change model** with data as an input and economies of scope
1. Empirically study a prototypical data-intensive sector with significant state interest: facial recognition AI in China

2. Directed technical change model with data as an input and economies of scope

3. Applications illustrating varied ways that data-intensive innovation may be shaped by the state
   - Directly: industrial policy
   - Indirectly: surveillance levels; privacy regulation
Evidence from facial recognition AI firms in China
Government procurement and facial recognition AI in China

Prototypical setting in which to study the role of the state in data-intensive innovation

- Developing facial recognition software requires access to large datasets, in particular allowing linking faces to personal identifiers
- Chinese government units collect this type of data and demand AI software in order to surveil citizens
Government procurement and facial recognition AI in China

Prototypical sector/country in which to study the role of the state in data-intensive innovation

- Developing facial recognition software requires access to large datasets, in particular allowing linking faces to personal identifiers
- Chinese government units collect this type of data and demand AI software in order to surveil citizens

The mechanism

1. Consider a firm that receives a government procurement contract
   - e.g., facial recognition software for a municipal police department
Government procurement and facial recognition AI in China

Prototypical sector/country in which to study the role of the state in data-intensive innovation

- Developing facial recognition software requires access to large datasets, in particular allowing linking faces to personal identifiers
- Chinese government units collect this type of data and demand AI software in order to surveil citizens

The mechanism

1. Consider a firm that receives a government procurement contract
   - e.g., facial recognition software for a municipal police department
2. It receives access to government data which is not publicly available
   - e.g., video from street cameras and labeled personal images
Government procurement and facial recognition AI in China

Prototypical sector/country in which to study the role of the state in data-intensive innovation

- Developing facial recognition software requires access to large datasets, in particular allowing linking faces to personal identifiers
- Chinese government units collect this type of data and demand AI software in order to surveil citizens

The mechanism

1. Consider a firm that receives a government procurement contract
   - e.g., facial recognition software for a municipal police department

2. It receives access to government data which is not publicly available
   - e.g., video from street cameras and labeled personal images

3. It uses this data to develop government software, but also commercial
   - e.g., trains an algorithm for detecting individuals in video from street cameras, and uses that same data to train another algorithm to detect shoppers in video from retail stores’ cameras
Prototypical sector/country in which to study the role of the state in data-intensive innovation

- Developing facial recognition software requires access to large datasets, in particular allowing linking faces to personal identifiers
- Chinese government units collect this type of data and demand AI software in order to surveil citizens

The mechanism

1. Consider a firm that receives a government procurement contract
2. It receives access to government data which is not publicly available
3. It uses this data to develop government software, but also commercial

Note: economies of scope may not arise even if data is sharable, e.g., when govt soft crowds out resources from commercial soft
Empirical challenges

Data

1. Dataset linking AI firms to govt contracts did not exist

2. Dataset on AI firms software did not exist. Also, critical for us to classify by use (e.g., govt vs commercial)

3. No available direct measures of firm-level use of govt data
Empirical challenges

Data

1. Dataset linking AI firms to govt contracts did not exist
2. Dataset on AI firms software did not exist. Also, critical for us to classify by use (e.g., govt vs commercial)
3. No available direct measures of firm-level use of govt data

Identification

1. Non-random assignment of govt contracts
2. Contracts work through other mechanisms unrelated to data
Data 1. Linking AI firms to gov’t. contracts

1. Identify all facial recognition AI firms
   - Two sources: Tianyancha (People’s Bank of China) and PitchBook (Morningstar)

2. Obtain universe of government contracts
   - Source: Chinese Govt. Procurement Database (Ministry of Finance)
Data 1. Linking AI firms to gov’t. contracts

1. Identify all facial recognition AI firms
   - Two sources: Tianyancha (People’s Bank of China) and PitchBook (Morningstar)

2. Obtain universe of government contracts
   - Source: Chinese Govt. Procurement Database (Ministry of Finance)

3. Link Government Buyers to AI Suppliers
Data 2. AI firms’ software

Registered with Min. of Industry and Information Technology
- Validation exercise: check against IPO Prospectus of MegVii

Categorize by intended customers (via ML text analysis):

1. **Government**: e.g., *smart city — real time monitoring system on main traffic routes*;
2. **Commercial**: e.g., *visual recognition system for smart retail*;
3. **General**: e.g., *A synchronization method for multi-view cameras based on FPGA chips.*
Data 3. Measuring access to government data

**Variation across contract types:** public security contracts “richer in data” than non-public security (e.g., police v. banks)

- Contract with police department in Heilongjiang Province on August 29th, 2018: AI firm shall provide a facial recognition system that can store and analyze at least 30 million facial images

- Contract with provincial bank in Gansu Province on November 20th, 2018: system to provide identification services for the bank’s clients
Data 3. Measuring access to government data

**Variation across contract types:** public security contracts “richer in data” than non-public security (e.g., police v. banks)

- Contract with police department in Heilongjiang Province on August 29th, 2018: AI firm shall provide a facial recognition system that can store and analyze at least 30 million facial images

- Contract with provincial bank in Gansu Province on November 20th, 2018: system to provide identification services for the bank’s clients

**Variation within AI public security contracts:** prefectures with more/fewer advanced street cameras

1. Identify non-AI contracts: police department purchases of street cameras
2. Measure quantity of advanced cameras in a prefecture at a given time
3. Categorize public security contracts as coming from “high” or “low” camera capacity prefectures
Baseline empirical strategy

- **Today**: results within public security contracts (see paper for public v. non-public)

- **Triple diffs**: compare cumulative software releases before and after firms received 1st data-rich contracts, relative to the data-scarce ones

\[ y_{it} = \sum_{T} \beta_1 T T_{it} Data_i + \sum_{T} \beta_2 T T_{it} + \alpha_t + \gamma_i + \sum_{T} T_{it} X_i + \epsilon_{it} \]

- \( T_{it} \): 1 if, at time \( t \), \( T \) semi-years have passed before/since firm \( i \) received 1st contract

- \( Data_i \): 1 if firm \( i \) receives contract from “high” camera capacity prefecture (measured at time of contract receipt)

- \( X_i \) controls for pre-contract firm characteristics: age, size (cap), and software production
Public security contract “richer in data” \(\Rightarrow\) Larger increase in both government AND commercial software

(a) Government

(b) Commercial
Additional evidence on the importance of data as an input

1. **Bids** are lower for data-rich contracts (even controlling for firm fixed effects). Also, more **bidders** overall.

2. **Data-complementary** software (e.g., storage/transmission) differentially increases when obtaining a data-rich contract.

3. Look at firms producing **video facial recognition**, which is particularly data-intensive.
   - Coefficients almost double in magnitude compared to baseline.
Evaluating alternative hypotheses

1. **Selection** at a given time differs by contract
   - Firm controls. No differential pre-contract levels/trends of software

2. **Terms and tasks** differ by contract
   - Descriptions of data-rich and -scarce contracts are similar in content
   - Similar govt soft produced after data-rich and -scarce contracts too

3. **Learning by doing** differs by contract
   - Previous result on description of govt software suggests otherwise
   - Control for pre-contract govt software, or in own/opposite category

4. **Importance of capital** differs by contract
   - Control for time-period×: pre-contract market cap or amount of external financing, and monetary value of contract

5. **Signals** differ by contract
   - Subsamples of firms: (i) from a *mother* firm that has already received contract, or (ii) receiving a 2nd data-rich contract

6. **Govt connections** or **opportunities** differ by contract
   - Drop contracts with Beijing/Shangai or firm’s home province.
   - Control for time-period×GDP-per-cap. “Placebo” with non-public security contracts from prefectures with above-median surveillance
Macroeconomic implications of firms’ access to government data
Macro implications of firms’ access to government data

- **Micro evidence**: $\uparrow$ govt data $\implies$ $\uparrow$ firm innovation

- Does not imply: $\uparrow$ govt data $\implies$ $\uparrow$ aggregate innovation

- E.g., crowd out of resources from other firms

- Nor that such policy would be desirable even if it did

- Innovation crowds out resources from consumption. Welfare?

- Preferences for surveillance and data collection?
Macro implications of firms’ access to government data

- **Micro evidence:** $\uparrow$ govt data $\implies\uparrow$ firm innovation
- **Does not imply:** $\uparrow$ govt data $\implies\uparrow$ aggregate innovation
  - E.g., crowd out of resources from other firms

Welfare? Preferences for surveillance and data collection?
Macro implications of firms’ access to government data

- Micro evidence: ↑ govt data $\implies$ ↑ firm innovation

- Does not imply: ↑ govt data $\implies$ ↑ aggregate innovation
  - E.g., crowd out of resources from other firms

- Nor that such policy would be desirable even if it did
  - Innovation crowds out resources from consumption. Welfare?
  - Preferences for surveillance and data collection?
Model (Sketch)

- 3 intermediate sectors: **Commercial** software; **Government** software; **Non**-software
- Commercial soft and Non-soft combined to produce **final good** (CES). Consumed by **rep. household**
- **State** uses govt software to produce **surveillance**
Model (Sketch)

- 3 intermediate sectors: **Commercial** software; **Government** software; **Non**-software

- Commercial soft and Non-soft combined to produce **final good** (CES). Consumed by **rep. household**

- **State** uses govt software to produce **surveillance**

- **Innovators** develop and supply differentiated varieties of each type of intermediate

- **Technologies** for variety production
  - Gov soft uses **govt data** $d_g$ and goods $x_g$ as inputs: $F_g(d_g, x_g)$
Model (Sketch)

- 3 intermediate sectors: Commercial software; Government software; Non-software
- Commercial soft and Non-soft combined to produce final good (CES). Consumed by rep. household
- State uses govt software to produce surveillance
- Innovators develop and supply differentiated varieties of each type of intermediate
- Technologies for variety production
  - Gov soft uses govt data \( d_g \) and goods \( x_g \) as inputs: \( F_g(\ d_g , x_g ) \)
  - Economies of scope: same \( d_g \) is an input into Comm soft
  - Comm soft also uses private data \( d_p \): \( F_c(\ d_g , d_p , x_p ) \)
Model (Sketch)

- 3 intermediate sectors: **Commercial** software; **Government** software; **Non**-software

- Commercial soft and Non-soft combined to produce **final good** (CES). Consumed by **rep. household**

- **State** uses govt software to produce **surveillance**

- **Innovators** develop and supply differentiated varieties of each type of intermediate

- **Technologies** for variety production
  - Gov soft uses **govt data** $d_g$ and goods $x_g$ as inputs: $F_g(d_g, x_g)$
  - **Economies of scope:** same $d_g$ is an input into Comm soft
  - Comm soft also uses **private data** $d_p$: $F_c(d_g, d_p, x_p)$
  - Assume $d_g$ and $d_p$ are **gross substitutes**
Model (Sketch)

- 3 intermediate sectors: **Commercial** software; **Government** software; **Non**-software

- Commercial soft and Non-soft combined to produce **final good** (CES). Consumed by **rep. household**

- **State** uses govt software to produce **surveillance**

- **Innovators** develop and supply differentiated varieties of each type of intermediate

- **Technologies** for variety production

  - Gov soft uses **govt data** $d_g$ and goods $x_g$ as inputs: $F_g(d_g, x_g)$
  - **Economies of scope:** same $d_g$ is an input into Comm soft
  - Comm soft also uses **private data** $d_p$: $F_c(d_g, d_p, x_p)$
  - Assume $d_g$ and $d_p$ are **gross substitutes**
  - Non soft does not use data, only goods: $F_z(x_z)$
Firms can purchase private data in the market

As in our empirical context, no market for govt data $d_g$. Can only be accessed by obtaining a contract to produce govt varieties:

- Pay fixed cost $F$ and obtain govt contract with prob $\lambda$. Obtain fixed quantity of govt data $\bar{d}_g$
Model (Sketch cont.)

- Firms can purchase **private data in the market**

- As in our empirical context, **no market for govt data** $d_g$. Can only be **accessed by obtaining a contract** to produce govt varieties:
  
  - Pay fixed cost $F$ and obtain govt contract with prob $\lambda$. Obtain **fixed quantity of govt data** $\bar{d}_g$

- **Innovation** on new varieties as in *Acemoglu (2002)*
  
  - 1 unit of final good produces $\mu_i$ new varieties in sector $i$. Innovator becomes monopolist of those varieties
Firms can purchase **private data in the market**

As in our empirical context, **no market for govt data** $d_g$. Can only be **accessed by obtaining a contract** to produce govt varieties:

- Pay fixed cost $F$ and obtain govt contract with prob $\lambda$. Obtain fixed quantity of govt data $\bar{d}_g$

**Innovation** on new varieties as in Acemoglu (2002)

- 1 unit of final good produces $\mu_i$ new varieties in sector $i$. Innovator becomes monopolist of those varieties

**Govt data supply:** a **by-product** of the state’s **surveillance**

**Private data supply:** a **by-product** of total private transactions (measured by final good output)
BGP equilibrium

- **State policy:** govt data provision to firms $\tilde{d}_g$, surveillance spending and lump-sum taxes such that budget is balanced.

- **Existence/uniqueness of BGP** with free-entry of innovators and **three types of firms** producing:
  1. both govt and comm soft, using govt and private data
  2. only comm soft, using private data
  3. non-soft
Macroeconomic implications of $\uparrow \bar{d}_g$

**Theorem**
- Assume comm soft demand is sufficiently elastic and conditions for a unique BGP to exist are satisfied
- $\uparrow \bar{d}_g \iff \uparrow \text{growth rate and bias of private innovation towards data-intensive software}$
Macroeconomic implications of $\uparrow \bar{d}_g$

**Theorem**
- Assume comm soft demand is sufficiently elastic and conditions for a unique BGP to exist are satisfied
- $\uparrow \bar{d}_g \Rightarrow \uparrow$ growth rate and bias of private innovation towards data-intensive software

**Comments**
1. Choose $\bar{d}_g \Rightarrow$ surveillance spending endogenous, or, choose surveillance spending $\Rightarrow \bar{d}_g$ endogenous
Macroeconomic implications of \( \uparrow \bar{d}_g \)

**Theorem**
- Assume comm soft demand is sufficiently elastic and conditions for a unique BGP to exist are satisfied
- \( \uparrow \bar{d}_g \implies \uparrow \text{growth} \) rate and **bias** of private innovation towards data-intensive software

**Comments**
1. Choose \( \bar{d}_g \Rightarrow \) surveillance spending endogenous, or, choose surveillance spending \( \Rightarrow \bar{d}_g \) endogenous

2. Benevolent state should **maximize welfare**, not growth \( \eta \)

\[
U = \frac{1}{\rho - \eta(1 - \theta)} \frac{(C/Y)^{1-\theta}}{1 - \theta}
\]

- **Two offsetting forces** that decrease \( C/Y \)
  - Innovation crowds out final good from consumption
  - Surveillance production does too
Second-best government data $\bar{d}_g$ provision

Remark 1: reminiscent of Barro (1990) studying provision of govt goods/services that increase firm productivity (e.g. infrastructure)
Remark 1: reminiscent of Barro (1990) studying provision of govt goods/services that increase firm productivity (e.g. infrastructure)
Second-best government data $\bar{d}_g$ provision

Remark 2: interior solution for our baseline parameterization ($\alpha = 0.82$). But, if economies of scope are relatively weak, then government data provision (and surveillance) is not desirable.
Contributions and Takeaways from Applications

1. To the literature on the economics of AI (e.g., Aghion et al., 2017; Agrawal et al., 2018; Farboodi et al., 2019; Jones and Tonetti, 2019)

   - We highlight role of the state as collector and repository of data, and economies of scope from sharability of data across uses within the firm

2. To the literature on industrial and innovation policies (e.g., Rodrik, 2007; Lane, 2020; Bloom et al., 2019)

   - We study a frontier technology and propose a mechanism through which governments affect innovation

   - Application on industrial policy: government data provision is justified for reasons that differ from those of traditional industrial policies

3. To the literature on the political economy of innovation

   - Application on surveillance states: objectives actually align with data-intensive innovation (in contrast with e.g., North, 1991; Acemoglu and Robinson, 2006, 2012), but may still be detrimental to citizens

   - Application on privacy regulation: interaction between norms, policies, and data-intensive innovation (similar to Benabou, 2015; and Besley and Persson, 2019)
Contributions and Takeaways from Applications

1. To the literature on the economics of AI (e.g., Aghion et al., 2017; Agrawal et al., 2018; Farboodi et al., 2019; Jones and Tonetti, 2019)
   - We highlight role of the state as collector and repository of data, and economies of scope from sharability of data across uses within the firm.

2. To the literature on industrial and innovation policies (e.g., Rodrik, 2007; Lane, 2020; Bloom et al., 2019)
   - We study a frontier technology and propose a mechanism through which governments affect innovation.
   - Application on industrial policy: government data provision is justified for reasons that differ from those of traditional industrial policies.
Contributions and Takeaways from Applications

1. To the literature on the economics of AI (e.g., Aghion et al., 2017; Agrawal et al., 2018; Farboodi et al., 2019; Jones and Tonetti, 2019)
   - We highlight role of the state as collector and repository of data, and economies of scope from sharability of data across uses within the firm

2. To the literature on industrial and innovation policies (e.g., Rodrik, 2007; Lane, 2020; Bloom et al., 2019)
   - We study a frontier technology and propose a mechanism through which governments affect innovation
   - Application on industrial policy: government data provision is justified for reasons that differ from those of traditional industrial policies

3. To the literature on the political economy of innovation
   - Application on surveillance states: objectives actually align with data-intensive innovation (in contrast with e.g., North, 1991; Acemoglu and Robinson, 2006, 2012), but may still be detrimental to citizens
Contributions and Takeaways from Applications

1. To the literature on the economics of AI (e.g., Aghion et al., 2017; Agrawal et al., 2018; Farboodi et al., 2019; Jones and Tonetti, 2019)
   - We highlight role of the state as collector and repository of data, and economies of scope from sharability of data across uses within the firm

2. To the literature on industrial and innovation policies (e.g., Rodrik, 2007; Lane, 2020; Bloom et al., 2019)
   - We study a frontier technology and propose a mechanism through which governments affect innovation
   - Application on industrial policy: government data provision is justified for reasons that differ from those of traditional industrial policies

3. To the literature on the political economy of innovation
   - Application on surveillance states: objectives actually align with data-intensive innovation (in contrast with e.g., North, 1991; Acemoglu and Robinson, 2006, 2012), but may still be detrimental to citizens
   - Application on privacy regulation: interaction between norms, policies, and data-intensive innovation (similar to Benabou, 2015; and Besley and Persson, 2019)
Appendix
Categorization: analyze text using machine learning

- Frontier method: Recurrent Neural Network (RNN) model using tensorflow
  - Corpus: 13,000 manually labeled software programs
  - Word-embedding: converted sentences to vectors based on word frequencies and used the words from full datasets as dictionary
  - Long Short-Term Memory (LSTM) algorithm: 2 layers of 32 nodes
  - 90% of corpus for training, 10% for validating
  - 10,000 training cycles are run for gradient descent on loss function

- Results robust to perturbing parameters of learning model
Directed technical change model with data as an input

Key ingredients

- The state collects govt data and demands data-intensive government software to produce “surveillance services”

- Firms can access fixed quantity of govt data ($d_g$) by selling software to the state

- Economies of scope: govt data can be used to produce commercial software (input into final consumption good)

- Private data available in the market (gross sub with govt data)
Directed technical change model with data as an input

**Key ingredients**
- The state collects govt data and demands data-intensive government software to produce “surveillance services”
- Firms can access fixed quantity of govt data ($d_g$) by selling software to the state
- Economies of scope: govt data can be used to produce commercial software (input into final consumption good)
- Private data available in the market (gross sub with govt data)

**BGP** with free-entry of innovators and **three types of firms**:  
1. both govt and comm soft, using govt and private data  
2. only comm soft, using private data  
3. non-software
Chinese citizens value data protection significantly less

**Putting a Price on Data**

Surveys of consumers in the United States, China, India, Great Britain, and Germany reveal that they value some types of information much more highly than others. The approximate amount people say they would pay to protect each data type (per person, US$, 2014):
Government privacy protections may limit data collection

San Francisco is first US city to ban facial recognition

By Dave Lee
North America technology reporter

15 May 2019
(a) Bid price (bin scatter)

(b) # of Bidders (bin scatter)
(a) Data-complementary

(b) Government (data-comp as proxy)

(c) Commercial (data-comp as proxy)
(a) Government (for video-AI)

(b) Commercial (for video-AI)

(c) Data-complementary (for video-AI)
Table A.11: Scale effects and learning-by-doing

<table>
<thead>
<tr>
<th></th>
<th>Government</th>
<th>Commercial</th>
<th>Data-complementary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td><strong>Panel A: Baseline</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Semiyears Before</td>
<td>-0.177</td>
<td>-0.239</td>
<td>-0.310</td>
</tr>
<tr>
<td></td>
<td>(0.268)</td>
<td>(0.231)</td>
<td>(0.270)</td>
</tr>
<tr>
<td>6 Semiyears After</td>
<td>5.595***</td>
<td>5.811***</td>
<td>6.383***</td>
</tr>
<tr>
<td></td>
<td>(0.444)</td>
<td>(0.378)</td>
<td>(0.443)</td>
</tr>
<tr>
<td>4 Semiyears Before × High Capacity</td>
<td>-0.279</td>
<td>0.633</td>
<td>0.130</td>
</tr>
<tr>
<td></td>
<td>(0.620)</td>
<td>(0.539)</td>
<td>(0.627)</td>
</tr>
<tr>
<td>6 Semiyears After × High Capacity</td>
<td>2.911***</td>
<td>1.861***</td>
<td>2.766***</td>
</tr>
<tr>
<td></td>
<td>(0.642)</td>
<td>(0.550)</td>
<td>(0.644)</td>
</tr>
<tr>
<td><strong>Panel B: Control for government pre-contract software production</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Semiyears Before</td>
<td>0.138</td>
<td>-0.076</td>
<td>-0.081</td>
</tr>
<tr>
<td></td>
<td>(0.233)</td>
<td>(0.220)</td>
<td>(0.252)</td>
</tr>
<tr>
<td>6 Semiyears After</td>
<td>1.769***</td>
<td>3.846***</td>
<td>3.652***</td>
</tr>
<tr>
<td></td>
<td>(0.386)</td>
<td>(0.362)</td>
<td>(0.415)</td>
</tr>
<tr>
<td>4 Semiyears Before × High Capacity</td>
<td>0.170</td>
<td>0.869*</td>
<td>0.489</td>
</tr>
<tr>
<td></td>
<td>(0.538)</td>
<td>(0.514)</td>
<td>(0.586)</td>
</tr>
<tr>
<td>6 Semiyears After × High Capacity</td>
<td>1.477***</td>
<td>1.116**</td>
<td>1.722***</td>
</tr>
<tr>
<td></td>
<td>(0.556)</td>
<td>(0.525)</td>
<td>(0.602)</td>
</tr>
<tr>
<td><strong>Panel C: Control for same category pre-contract software production</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Semiyears Before</td>
<td>0.138</td>
<td>0.034</td>
<td>-0.047</td>
</tr>
<tr>
<td></td>
<td>(0.233)</td>
<td>(0.209)</td>
<td>(0.253)</td>
</tr>
<tr>
<td>6 Semiyears After</td>
<td>1.769***</td>
<td>2.577***</td>
<td>3.173***</td>
</tr>
<tr>
<td></td>
<td>(0.386)</td>
<td>(0.344)</td>
<td>(0.418)</td>
</tr>
<tr>
<td>4 Semiyears Before × High Capacity</td>
<td>0.170</td>
<td>0.841*</td>
<td>0.361</td>
</tr>
<tr>
<td></td>
<td>(0.538)</td>
<td>(0.487)</td>
<td>(0.589)</td>
</tr>
<tr>
<td>6 Semiyears After × High Capacity</td>
<td>1.477***</td>
<td>1.132**</td>
<td>2.013***</td>
</tr>
<tr>
<td></td>
<td>(0.556)</td>
<td>(0.498)</td>
<td>(0.605)</td>
</tr>
<tr>
<td><strong>Panel D: Control for opposite category pre-contract software production</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Semiyears Before</td>
<td>0.080</td>
<td>-0.076</td>
<td>-0.061</td>
</tr>
<tr>
<td></td>
<td>(0.250)</td>
<td>(0.220)</td>
<td>(0.256)</td>
</tr>
<tr>
<td>6 Semiyears After</td>
<td>2.399***</td>
<td>3.846***</td>
<td>3.474***</td>
</tr>
<tr>
<td></td>
<td>(0.416)</td>
<td>(0.362)</td>
<td>(0.423)</td>
</tr>
<tr>
<td>4 Semiyears Before × High Capacity</td>
<td>-0.078</td>
<td>0.869*</td>
<td>0.302</td>
</tr>
<tr>
<td></td>
<td>(0.579)</td>
<td>(0.514)</td>
<td>(0.596)</td>
</tr>
<tr>
<td>6 Semiyears After × High Capacity</td>
<td>2.231***</td>
<td>1.116**</td>
<td>2.111***</td>
</tr>
<tr>
<td></td>
<td>(0.599)</td>
<td>(0.525)</td>
<td>(0.612)</td>
</tr>
<tr>
<td></td>
<td>Government</td>
<td>Commercial</td>
<td>Data-complementary</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>------------</td>
<td>------------</td>
<td>--------------------</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td><strong>Panel A: Baseline</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Semiyears Before</td>
<td>-0.177</td>
<td>-0.239</td>
<td>-0.310</td>
</tr>
<tr>
<td></td>
<td>(0.268)</td>
<td>(0.231)</td>
<td>(0.270)</td>
</tr>
<tr>
<td>6 Semiyears After</td>
<td>5.595***</td>
<td>5.811***</td>
<td>6.383***</td>
</tr>
<tr>
<td></td>
<td>(0.444)</td>
<td>(0.378)</td>
<td>(0.443)</td>
</tr>
<tr>
<td>4 Semiyears Before × High Capacity</td>
<td>-0.279</td>
<td>0.633</td>
<td>0.130</td>
</tr>
<tr>
<td></td>
<td>(0.620)</td>
<td>(0.539)</td>
<td>(0.627)</td>
</tr>
<tr>
<td>6 Semiyears After × High Capacity</td>
<td>2.911***</td>
<td>1.861***</td>
<td>2.766***</td>
</tr>
<tr>
<td></td>
<td>(0.642)</td>
<td>(0.550)</td>
<td>(0.644)</td>
</tr>
<tr>
<td><strong>Panel B: Sample — not first contract within mother firm</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Semiyears Before</td>
<td>-0.078</td>
<td>-0.431</td>
<td>-0.184</td>
</tr>
<tr>
<td></td>
<td>(0.213)</td>
<td>(0.362)</td>
<td>(0.283)</td>
</tr>
<tr>
<td>6 Semiyears After</td>
<td>4.606***</td>
<td>6.730***</td>
<td>6.370***</td>
</tr>
<tr>
<td></td>
<td>(0.332)</td>
<td>(0.557)</td>
<td>(0.438)</td>
</tr>
<tr>
<td>4 Semiyears Before × High Capacity</td>
<td>1.035</td>
<td>1.047</td>
<td>0.820</td>
</tr>
<tr>
<td></td>
<td>(0.786)</td>
<td>(1.384)</td>
<td>(1.081)</td>
</tr>
<tr>
<td>6 Semiyears After × High Capacity</td>
<td>2.753***</td>
<td>1.975*</td>
<td>1.024</td>
</tr>
<tr>
<td></td>
<td>(0.710)</td>
<td>(1.200)</td>
<td>(0.947)</td>
</tr>
<tr>
<td><strong>Panel C: Sample — second contract within subsidiary firm</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Semiyears Before</td>
<td>-1.577*</td>
<td>2.214***</td>
<td>2.015***</td>
</tr>
<tr>
<td></td>
<td>(0.916)</td>
<td>(0.656)</td>
<td>(0.697)</td>
</tr>
<tr>
<td>6 Semiyears After</td>
<td>8.533***</td>
<td>7.856***</td>
<td>13.538***</td>
</tr>
<tr>
<td></td>
<td>(1.430)</td>
<td>(1.025)</td>
<td>(1.088)</td>
</tr>
<tr>
<td>4 Semiyears Before × High Capacity</td>
<td>1.090</td>
<td>-1.943**</td>
<td>-1.819*</td>
</tr>
<tr>
<td></td>
<td>(1.287)</td>
<td>(0.923)</td>
<td>(0.980)</td>
</tr>
<tr>
<td>6 Semiyears After × High Capacity</td>
<td>29.042***</td>
<td>2.876**</td>
<td>17.833***</td>
</tr>
<tr>
<td></td>
<td>(1.881)</td>
<td>(1.349)</td>
<td>(1.432)</td>
</tr>
</tbody>
</table>
### Table A.13: Robustness — firm geography

<table>
<thead>
<tr>
<th></th>
<th>Government</th>
<th>Commercial</th>
<th>Data-complementary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>Panel A: Baseline</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Semiyears Before</td>
<td>-0.177</td>
<td>-0.239</td>
<td>-0.310</td>
</tr>
<tr>
<td></td>
<td>(0.268)</td>
<td>(0.231)</td>
<td>(0.270)</td>
</tr>
<tr>
<td>6 Semiyears After</td>
<td>5.595***</td>
<td>5.811***</td>
<td>6.383***</td>
</tr>
<tr>
<td></td>
<td>(0.444)</td>
<td>(0.378)</td>
<td>(0.443)</td>
</tr>
<tr>
<td>4 Semiyears Before × High Capacity</td>
<td>-0.279</td>
<td>0.633</td>
<td>0.130</td>
</tr>
<tr>
<td></td>
<td>(0.620)</td>
<td>(0.539)</td>
<td>(0.627)</td>
</tr>
<tr>
<td>6 Semiyears After × High Capacity</td>
<td>2.911***</td>
<td>1.861***</td>
<td>2.766***</td>
</tr>
<tr>
<td></td>
<td>(0.642)</td>
<td>(0.550)</td>
<td>(0.644)</td>
</tr>
<tr>
<td>Panel B: Drop Beijing, Shanghai</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Semiyears Before</td>
<td>-0.179</td>
<td>-0.242</td>
<td>-0.277</td>
</tr>
<tr>
<td></td>
<td>(0.264)</td>
<td>(0.166)</td>
<td>(0.249)</td>
</tr>
<tr>
<td>6 Semiyears After</td>
<td>5.511***</td>
<td>5.873***</td>
<td>6.286***</td>
</tr>
<tr>
<td></td>
<td>(0.423)</td>
<td>(0.264)</td>
<td>(0.397)</td>
</tr>
<tr>
<td>4 Semiyears Before × High Capacity</td>
<td>-0.114</td>
<td>0.763*</td>
<td>0.235</td>
</tr>
<tr>
<td></td>
<td>(0.634)</td>
<td>(0.404)</td>
<td>(0.603)</td>
</tr>
<tr>
<td>6 Semiyears After × High Capacity</td>
<td>2.983***</td>
<td>1.118***</td>
<td>2.863***</td>
</tr>
<tr>
<td></td>
<td>(0.641)</td>
<td>(0.403)</td>
<td>(0.605)</td>
</tr>
<tr>
<td>Panel C: Firm based outside contract province</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Semiyears Before</td>
<td>-0.195</td>
<td>-0.165</td>
<td>-0.293</td>
</tr>
<tr>
<td></td>
<td>(0.209)</td>
<td>(0.245)</td>
<td>(0.218)</td>
</tr>
<tr>
<td>6 Semiyears After</td>
<td>5.254***</td>
<td>5.862***</td>
<td>6.153***</td>
</tr>
<tr>
<td></td>
<td>(0.333)</td>
<td>(0.387)</td>
<td>(0.346)</td>
</tr>
<tr>
<td>4 Semiyears Before × High Capacity</td>
<td>-0.053</td>
<td>0.721</td>
<td>0.177</td>
</tr>
<tr>
<td></td>
<td>(0.555)</td>
<td>(0.658)</td>
<td>(0.586)</td>
</tr>
<tr>
<td>6 Semiyears After × High Capacity</td>
<td>2.365***</td>
<td>2.747***</td>
<td>2.815***</td>
</tr>
<tr>
<td></td>
<td>(0.542)</td>
<td>(0.636)</td>
<td>(0.567)</td>
</tr>
</tbody>
</table>