Data, AI, and the State: Evidence from China

Martin Beraja (MIT)

NBER Digitization Tutorial, Spring 2022

The interplay between data, AI, and the state

- Data-intensive technologies (like AI) can transform modern economies but have brought **new challenges** to the fore
- This has raised questions about **the role of governments**
  1. When technologies use data, what innovation policies and regulations are appropriate?
  2. Could governments misuse AI as a tool of repression and social control?

The interplay between data, AI, and the state

- Data-intensive technologies (like AI) can transform modern economies but have brought **new challenges** to the fore
- This has raised questions about **the role of governments**
  1. When technologies use data, what innovation policies and regulations are appropriate?
  2. Could governments misuse AI as a tool of repression and social control?
- **Today:** think about the **interplay** between data, AI, and the state
  - Emphasize that AI is not only data-intensive but also **dual-use**
  - **Prototypical** setting to study this question: **the facial recognition AI industry in China**

Outline

1. Data-intensive Innovation and the State: Evidence from AI firms in China (with David Yang and Noam Yuchtman)

2. AI-tocracy (with Andrew Kao, David Yang and Noam Yuchtman)

3. Exporting Autocracy via AI (with Andrew Kao, David Yang and Noam Yuchtman)

Motivation: government data as input in AI innovation

- AI innovation is **data-intensive**
  - Many recent AI advances made with decades-old algorithms applied to newly available big data

Motivation: government data as input in AI innovation

- AI innovation is **data-intensive**
  - Many recent AI advances made with decades-old algorithms applied to newly available big data
- Literature has focused on how data collected by **private** firms shapes AI innovation
  (Agrawal et al., 2019; Jones and Tonetti, 2020)

Motivation: government data as input in AI innovation

- AI innovation is **data-intensive**
  - Many recent AI advances made with decades-old algorithms applied to newly available big data
- Literature has focused on how data collected by **private** firms shapes AI innovation (Agrawal et al., 2019; Jones and Tonetti, 2020)
- Yet, throughout history, **states** have also collected massive quantities of data (Scott, 1998)
- The state has a large role in many areas
  - Public security, health care, education, basic science...

Motivation: government data as input in AI innovation

- AI innovation is **data-intensive**
  - Many recent AI advances made with decades-old algorithms applied to newly available big data
- Literature has focused on how data collected by **private** firms shapes AI innovation (Agrawal et al., 2019; Jones and Tonetti, 2020)
- Yet, throughout history, **states** have also collected massive quantities of data (Scott, 1998)
- The state has a large role in many areas
  - Public security, health care, education, basic science...

$\Longrightarrow$ **Government data** can exceed privately-collected data in magnitude / scope; or lack good substitutes altogether

Motivation: China's facial recognition AI sector

- A common way in which AI firms **gain access** to valuable government data is by **providing services** to the state

Motivation: China's facial recognition AI sector

- A common way in which AI firms **gain access** to valuable government data is by **providing services** to the state
- Think about **facial recognition AI firms in China**...
  - Train algorithms with, e.g., video streams of faces from many angles
  - The state's public security units collect this form of data through their surveillance apparatus, and contract AI firms for services
  - AI firms gaining access to this data can use it to train algorithms and develop software

This paper

Does access to **government data** when providing AI services to the state stimulate **commercial** AI innovation?

This paper

Does access to **government data** when providing AI services to the state stimulate **commercial** AI innovation?

The mechanism(s)

1. If gov't data and algorithms are **sharable** across uses, they can be used to develop AI products for commercial markets
   (e.g., a facial recognition platform for retail stores)
2. Firms may **learn** to manage and utilize large datasets too

$\Longrightarrow$ a procurement contract with access to gov't data can fuel commercial innovation, overcoming **crowd-out** from the contract

Does access to **government data** when providing AI services to the state stimulate **commercial** AI innovation?

The mechanism(s)

1. If gov't data and algorithms are **sharable** across uses, they can be used to develop AI products for commercial markets
   (e.g., a facial recognition platform for retail stores)
2. Firms may **learn** to manage and utilize large datasets too

$\Longrightarrow$ a procurement contract with access to gov't data can fuel commercial innovation, overcoming **crowd-out** from the contract

Evidence of this in China's facial recognition AI sector

Data 1: linking AI firms to govt. contracts

1. Identify all facial recognition AI **firms**
    - 7,837 firms
    - Two sources: Tianyancha (People's Bank of China) and PitchBook (Morningstar)

Data 1: linking AI firms to govt. contracts

1. Identify all facial recognition AI **firms**
    - 7,837 firms
    - Two sources: Tianyancha (People's Bank of China) and PitchBook (Morningstar)

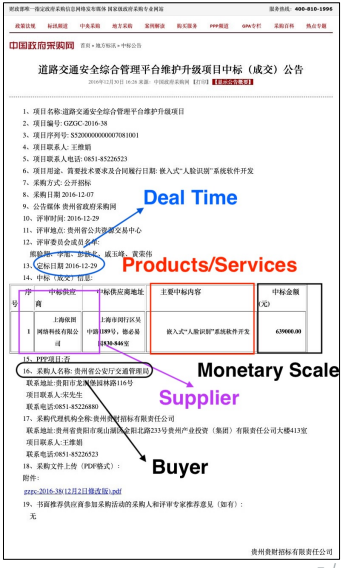
2. Obtain universe of **government contracts**
    - 2,997,105 contracts
    - Source: Chinese Govt. Procurement Database (Ministry of Finance)

Data 1: linking AI firms to govt. contracts

1. Identify all facial recognition AI **firms**
   - 7,837 firms
   - Two sources: Tianyancha (People's Bank of China) and PitchBook (Morningstar)

2. Obtain universe of **government contracts**
   - 2,997,105 contracts
   - Source: Chinese Govt. Procurement Database (Ministry of Finance)

3. Link government **buyers** to AI **suppliers**
   - 10,677 AI contracts issued by public security arms of government (e.g., local police department)
   - Data also on procurement of **AI-capable surveillance cameras**



Data 2: AI firms' software production

Registered with Min. of Industry and Information Technology

Categorize by intended customers (with RNN model using tensorflow):

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

Data 3: measuring access to government data

Within AI public security contracts: variation in the data collection capacity of the public security agency's local surveillance network

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

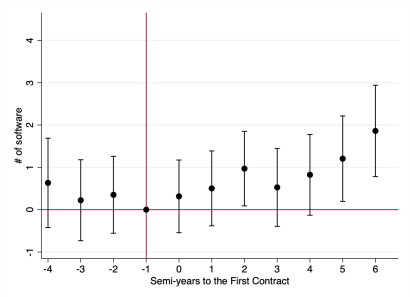
Baseline empirical strategy

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

$$y_{it} = \sum_T \beta_{1T} T_{it} Data_i + \sum_T \beta_{2T} T_{it} + \alpha_t + \gamma_i + \sum_T \beta_{3T} T_{it} X_i + \epsilon_{it}$$

  - 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 "data rich" contract (i.e., from "high" camera capacity prefecture at time of contract receipt)
  - X_i controls for pre-contract firm characteristics: age, size (cap), and software production

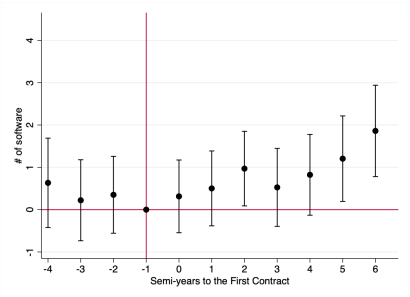
Commercial cumulative software releases



Magnitude: 2 new products over 3 years
(20% of pre-contract software)

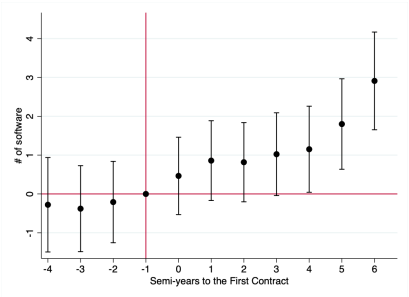
Public security contract "richer in data" & firm innovation

Commercial cumulative software releases



Magnitude: 2 new products over 3 years
(20% of pre-contract software)

Government cumulative software releases



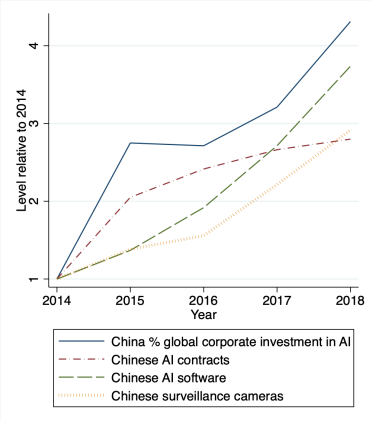
Commercial innov. overcomes gov't crowd-out

In the paper: assess mechanism(s) and alternative hypotheses

1. **Selection** at a given time differs by contract?
2. **Productive benefits other than data** differ by contract?
3. **Data/algorithm sharability** v. learning?

Takeaways

1. Access to gov't data contributed to Chinese AI firms' emergence as leading innovators
   - Indeed, this has coincided with the expansion of the state's AI procurement and surveillance capacity



Takeaways

1. Access to gov't data contributed to Chinese AI firms' emergence as leading innovators
   - Indeed, this has coincided with the expansion of the state's AI procurement and surveillance capacity

2. Novel role for the state in data-intensive economies
   - So far, emphasis on the regulation of privately-collected data due to antitrust or privacy concerns (Tirole, 2020; Aridor et al., 2020)
   - AI procurement and policies of gov't data collection and provision could, **whether intentionally or not**, stimulate and shape the direction of innovation in a range of sectors

Outline

1. Data-intensive Innovation and the State: Evidence from AI firms in China (with David Yang and Noam Yuchtman)

2. AI-tocracy (with Andrew Kao, David Yang and Noam Yuchtman)

3. Exporting Autocracy via AI (with Andrew Kao, David Yang and Noam Yuchtman)

Motivation: sustained innovation under autocracy?

Motivation: sustained innovation under autocracy?

- Conventional wisdom: autocracies are fundamentally misaligned with innovation
  (Lipset, 1959; Barro, 1996; Acemoglu and Robinson 2006; Glaeser et al., 2007; North et al. 2009)

Motivation: sustained innovation under autocracy?

- Conventional wisdom: autocracies are fundamentally misaligned with innovation
  (Lipset, 1959; Barro, 1996; Acemoglu and Robinson 2006; Glaeser et al., 2007; North et al. 2009)
- This paper: innovation in frontier techs. can be sustained under autocracy when they mutually reinforce each other

Motivation: sustained innovation under autocracy?

- **Conventional wisdom:** autocracies are fundamentally misaligned with innovation
  (Lipset, 1959; Barro, 1996; Acemoglu and Robinson 2006; Glaeser et al., 2007; North et al. 2009)
- **This paper:** innovation in frontier techs. can be sustained under autocracy when they **mutually reinforce each other**
  1. Frontier technology increases autocrats' probability of maintaining power
  2. Autocrats' spending on this tech. generates broader innovation spillovers

Motivation: sustained innovation under autocracy?

- Conventional wisdom: autocracies are fundamentally misaligned with innovation
  (Lipset, 1959; Barro, 1996; Acemoglu and Robinson 2006; Glaeser et al., 2007; North et al. 2009)
- This paper: innovation in frontier techs. can be sustained under autocracy when they **mutually reinforce each other**
  1. Frontier technology increases autocrats' probability of maintaining power
  2. Autocrats' spending on this tech. generates broader innovation spillovers
- AI may possess features that lead to a mutually reinforcing relationship

Motivation: sustained innovation under autocracy?

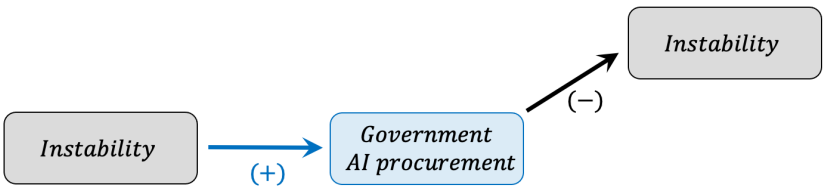
- Conventional wisdom: autocracies are fundamentally misaligned with innovation
  (Lipset, 1959; Barro, 1996; Acemoglu and Robinson 2006; Glaeser et al., 2007; North et al. 2009)
- This paper: innovation in frontier techs. can be sustained under autocracy when they **mutually reinforce each other**
  1. Frontier technology increases autocrats' probability of maintaining power
  2. Autocrats' spending on this tech. generates broader innovation spillovers
- AI may possess features that lead to a mutually reinforcing relationship
  1. As a technology of prediction, autocrats may be able to use AI for social / political control
     (Zuboff, 2019; Tirole, 2021; Acemoglu, 2021)
  2. Traditional spillovers (Moretti et al. 2019) + Sharability of gov't data/algorithms

AI-TOCRACY

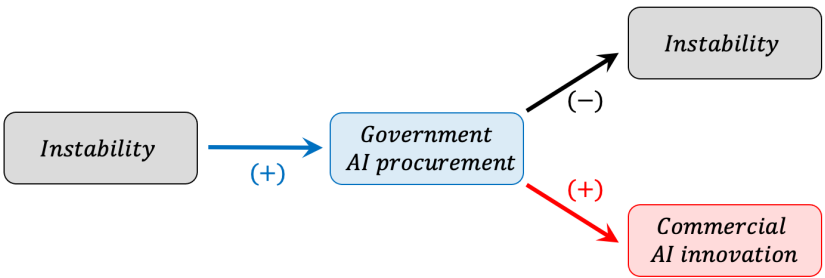
Test for a **mutually reinforcing relationship** between frontier innovation and autocracy in the context of China's facial recognition AI sector

Test for a **mutually reinforcing relationship** between frontier innovation and autocracy in the context of China's facial recognition AI sector

Test for a **mutually reinforcing relationship** between frontier innovation and autocracy in the context of China's facial recognition AI sector



Test for a **mutually reinforcing relationship** between frontier innovation and autocracy in the context of China's facial recognition AI sector



Data 4: local political unrest and weather

Protests and other episodes of political unrest:

- Daily level events in China from GDELT, a database tracking hundreds of news sites
- Use machine learning analysis to classify articles into those indicating political unrest (protests, demands, threats, etc.)
- There are 9,267 of these events from 2014 - 2020 throughout China

Data 4: local political unrest and weather

Protests and other episodes of political unrest:

- Daily level events in China from GDELT, a database tracking hundreds of news sites
- Use machine learning analysis to classify articles into those indicating political unrest (protests, demands, threats, etc.)
- There are 9,267 of these events from 2014 - 2020 throughout China

Weather

- Daily weather data from 260 weather stations across China
- LASSO regression to predict unrest events with 30 weather variables (e.g., temperature, precipitation, windspeed) and their interactions

↑ Local unrest in quarter $t \Longrightarrow$ ↑ public security AI procurement in $t+1$

1. **Diff-in-diff:** panel specification, controlling for location and time FEs
2. **IV:** instrument unrest with local weather conditions
3. **AI x Cameras:** complementarity?

	Public security AI procurement			
	(1)	(2)	(3)	(4)
Panel A.1: OLS, AI				
Unrest events	0.199***	0.198***	0.199***	0.200***
	(0.043)	(0.045)	(0.044)	(0.043)
Panel A.2: Lasso IV, AI				
Unrest events	0.388***	0.387***	0.388***	0.388***
	(0.088)	(0.088)	(0.088)	(0.087)
Panel B.1: OLS, AI X surveillance cameras				
Unrest events	0.681***	0.669***	0.680***	0.674***
	(0.154)	(0.157)	(0.155)	(0.150)
Panel B.2: Lasso IV, AI X surveillance cameras				
Unrest events	1.099***	1.083***	1.099***	1.085***
	(0.390)	(0.385)	(0.390)	(0.384)
GDP × time	Yes	No	No	Yes
Population × time	No	Yes	No	Yes
Gov. revenue × time	No	No	Yes	Yes

↑ Public security AI stock in quarter $t \Longrightarrow$ Effect on unrest $t+1$?

- Problematic to directly examine effect of AI stock on subsequent unrest events

  Positive autocorrelation between such events; and AI procurement is endogenous

↑ Public security AI stock in quarter $t \Longrightarrow \downarrow$ unrest due to good weather $t+1$

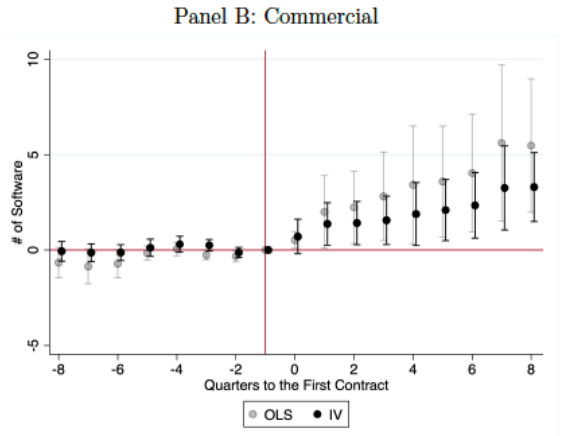
- Instead, examine whether AI tempers the effect of good weather on unrest events

  Also, look at AI in combinations with cameras, and placebo using non-public security AI

	Standardized number of unrest events							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Procurement of AI								
Favorable weather	0.9082***	0.9422***	0.9089***	0.9410***	0.9315***	0.9705***	0.9323***	0.9684***
	(0.1576)	(0.1564)	(0.1579)	(0.1510)	(0.1646)	(0.1632)	(0.1650)	(0.1574)
Public security procurement stock AI_{t-1}	-0.0096**	-0.0057	-0.0096**	-0.0044				
	(0.0048)	(0.0061)	(0.0048)	(0.0056)				
Favorable weather × public security AI_{t-1}	-0.2626*	-0.3152*	-0.2623*	-0.3088*				
	(0.1563)	(0.1742)	(0.1570)	(0.1687)				
Non-public security procurement stock AI_{t-1}					-0.0025	-0.0027	-0.0025	-0.0024
					(0.0017)	(0.0020)	(0.0017)	(0.0018)
Favorable weather × non-public security AI_{t-1}					-0.0492	-0.0576	-0.0495	-0.0535
					(0.0367)	(0.0411)	(0.0372)	(0.0375)
Panel B: Procurement of AI X procurement of surveillance cameras								
Favorable weather	0.8989***	0.9325***	0.8994***	0.9327***	0.9554***	0.9945***	0.9562***	0.9926***
	(0.1549)	(0.1524)	(0.1552)	(0.1480)	(0.1691)	(0.1659)	(0.1695)	(0.1605)
Public security procurement stock AI_{t-1}	0.2923***	0.3158***	0.2917***	0.3081***				
	(0.1083)	(0.0991)	(0.1081)	(0.0948)				
Favorable weather × public security AI_{t-1}	-0.7096***	-0.7952***	-0.7144***	-0.7789***				
	(0.2302)	(0.2412)	(0.2323)	(0.2248)				
Non-public security procurement stock AI_{t-1}					0.0605	0.0626	0.0608	0.0601
					(0.0600)	(0.0592)	(0.0603)	(0.0572)
Favorable weather × non-public security AI_{t-1}					0.7558	0.8049	0.7573	0.7744
					(0.6020)	(0.6015)	(0.6043)	(0.5801)
GDP × time	Yes	No	No	Yes	Yes	No	No	Yes
Log population × time	No	Yes	No	Yes	No	Yes	No	Yes
Gov. revenue × time	No	No	Yes	Yes	No	No	Yes	Yes

↑ Politically motivated public security AI procurement in quarter $t \Longrightarrow$ ↑ Commercial AI innovation in $t+1$

1. **Politically motivated public security contracts:** those from location with above median unrest at $t-1$
2. **Triple Diff:** before/after firms receive 1st politically motivated contract, then compare to non-public sec. contracts



Takeaways

1. Alignment between autocrats demand for social control and AI innovation
    - Could shed light on prominent episodes of frontier innovation under non-democracies
        - Aerospace technology in the USSR
        - Chemical engineering innovation in Imperial Germany

Takeaways

1. Alignment between autocrats demand for social control and AI innovation
   - Could shed light on prominent episodes of frontier innovation under non-democracies
     - Aerospace technology in the USSR
     - Chemical engineering innovation in Imperial Germany

2. If China exports autocracy-enhancing AI, what are the international ramifications?
   - China's comparative advantage in AI: $\uparrow$ state demand $\Longrightarrow$ $\uparrow$ firms' global competitiveness
   - Political bias: support autocracies and weak democracies abroad

Outline

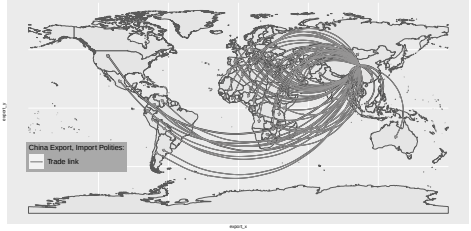
1. Data-intensive Innovation and the State: Evidence from AI firms in China (with David Yang and Noam Yuchtman)

2. AI-tocracy (with Andrew Kao, David Yang and Noam Yuchtman)

3. Exporting Autocracy via AI (with Andrew Kao, David Yang and Noam Yuchtman)

China's comparative advantage in AI

Exports more AI than the US, particularly so when compared to other frontier technologies

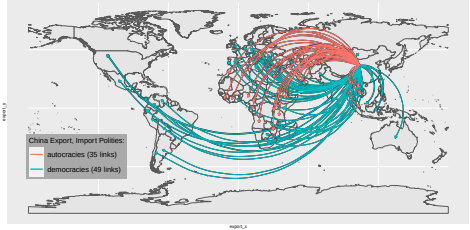


US vs. China, AI vs. frontier technologies

	Linear probability of trade			
	(1)	(2)	(3)	(4)
Origin China	-0.005	-0.005	0.005	-0.005
	(0.004)	(0.004)	(0.007)	(0.004)
AI	-2.948***	-2.913***	-3.053***	-2.849***
	(0.308)	(0.289)	(0.319)	(0.357)
Origin China X AI	0.220***	0.229***	0.262***	0.220***
	(0.045)	(0.045)	(0.061)	(0.045)
N	5364	5364	5364	5364

Notes: Regressions are at the product-import-export country dyad level. Outcome is dummy for trade. Omitted: US X not AI. Errors clustered at origin countries. All columns control for import/export GDP and log distance. Column (2) adds controls for common border, free trade agreements, and shared colonial background. Column (3) adds controls for common language, legal system, and religion. Column (4) adds controls for landlocked and island characteristics.

Imported more by autocracies and weak democracies. Not the case for the imports of China's other frontier technologies



AI vs. frontier technologies by polity type

	Linear probability of trade			
	(1)	(2)	(3)	(4)
Destination authoritarian	-0.043**	-0.043**	-0.046***	-0.041**
	(0.017)	(0.017)	(0.016)	(0.017)
Origin authoritarian	-0.107***	-0.107***	-0.110***	-0.105***
	(0.026)	(0.026)	(0.025)	(0.025)
Origin China	-0.058***	-0.059***	-0.048**	-0.054**
	(0.021)	(0.021)	(0.021)	(0.021)
AI	2.658***	2.657***	2.666***	2.668***
	(0.207)	(0.209)	(0.213)	(0.212)
Destination authoritarian X AI	0.046***	0.046***	0.049***	0.044***
	(0.017)	(0.017)	(0.016)	(0.017)
Origin authoritarian X AI	0.104***	0.104***	0.107***	0.102***
	(0.026)	(0.026)	(0.025)	(0.026)
Origin China X AI	0.545***	0.547***	0.536***	0.542***
	(0.030)	(0.033)	(0.031)	(0.031)
N	320796	320796	320796	320796

Notes: Regression at the product-import-export country dyad level. Outcome is dummy for trade. Omitted: origin/destination democracy X not AI. Errors two-way clustered at origin and destination countries. All columns control for import/export GDP and log distance. Column (2) adds controls for common border, free trade agreements, and shared colonial background. Column (3) adds controls for common language, legal system, and religion. Column (4) adds controls for landlocked and island characteristics.

Thank you!

And if you want to chat more...
maberaja@mit.edu