Original research article

BIG DATA & SOCIETY

Big Data & Society
April–June: 1–13

DOI: 10.1177/20539517251340600
journals.sagepub.com/home/bds
Sage

Global data empires: Analysing artificial intelligence data annotation in China and the USA

Tongyu Wu[1], James Muldoon[2] and Bingqing Xia[3]

Abstract

As the two leading countries in the development of artificial intelligence (AI) systems, China and the United States largely rely on separate AI infrastructure and data annotation ecosystems. Studies have focussed almost exclusively on data annotation associated with American and European companies, limiting our understanding of how this contrasts with the Chinese development of AI. This article provides a comparative analysis of the political economy of the Chinese and American/European AI data annotation ecosystems, focusing on the role of the state and the practice of outsourcing to data annotation institutions. It finds that while the US state plays a protectionist role concerning AI infrastructure such as semiconductors and data centres, it adopts a laissez-faire approach to data annotation. The Chinese state, however, understands it has a comparative advantage in data and invests heavily in its own data ecosystem while maintaining stringent regulations for Chinese tech companies. Secondly, many US companies outsource data annotation work to business process outsourcing centres and digital platforms, whereas Chinese companies maintain these activities in-house or, through a process of 'inland-sourcing', send this work to 'third-tier' cities in Chinese provinces to data labelling bases, often jointly managed by local government and private companies.

Keywords

Artificial intelligence, data annotation, AI data work, data production, platform labour, business process outsourcing

Introduction

The ongoing development of artificial intelligence (AI) has led to a rapidly growing data annotation industry. AI companies developing machine learning models require data to be curated, annotated and verified by human workers (Miceli et al., 2022a; Muldoon et al., 2024a, Shestakofsky, 2024). Current research of American and European firms contends that this work is often outsourced through digital labour platforms and business process outsourcing (BPO) companies to workers in the Global South (Miceli and Posada, 2022; Tubaro and Casilli, 2019; Tubaro et al., 2020). These workers are paid only a few cents per task on platforms or little more than one US dollar an hour at BPO centres, with limited employment protections and strict systems of discipline and surveillance (Gray and Suri, 2019). The demand for data annotation services continues to expand rapidly as billions of investment dollars flow towards AI development. The global data collection and labelling market size was estimated at $2.56 billion in 2024 and is expected to grow at a compound annual growth rate of 18% to reach over $13 billion by 2034 (Fact.MR 2023).

Data annotation is a global market with AI companies outsourcing work to digital platforms and BPO companies worldwide. However, the study of data annotation tends to concentrate on specific geographic regions influenced by existing contracts between AI and data annotation companies and the location and language skills of researchers. For example, this research includes France (and francophone Africa) (Casilli, 2025; Casilli and Tubaro, 2019; Le Ludec et al., 2023), Germany (Schmidt, 2019; Miceli et al., 2020); the USA (and Anglophone Africa) (Muldoon et al., 2023; Muldoon et al., 2024b); Eastern

[1]Department of Sociology, Zhejiang University, Sociology, Hangzhou, China
[2]Department of Sociology, University of Essex, Colchester UK
[3]School of Communication, East China Normal University, Shanghai, China

Corresponding author:
James Muldoon, Department of Sociology, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK.
Email: James.muldoon@essex.ac.uk

Europe (Miceli et al., 2020; Murgia, 2024), Latin America (Miceli and Posada, 2022; Posada, 2021), and India (Khanna and Chandran, 2023). A noticeable gap in this research is Chinese data annotation (Wu and Xia, 2023; Wu and Xia, 2025). One practical reason for this research gap is the timing of China's data annotation industry boom, which occurred in 2019 and coincided with the Covid pandemic. Pandemic restrictions enforced by the Chinese government significantly increased challenges for researchers in collecting empirical data (see Wang et al., 2022). The relative absence of empirical scholarship from China is problematic because it is the second largest investor in AI behind the USA and makes up a considerable portion of AI development (Maslej et al., 2024). China has announced an ambitious program for the development of its own AI infrastructure and foundation models and relies on its own ecosystem of data annotation (Knight, 2017). The lack of information about Chinese data annotation precludes an understanding of the significant differences between the US and Chinese contexts and risks universalising the experiences of US and European firms outsourcing to the Global South.

The article aims to enhance our understanding of the diversity of the data annotation industry and highlight the importance of national contextual factors through a comparative analysis of the political economy of American and Chinese AI data annotation ecosystems. It does so across two central lines of inquiry: AI companies' relationship with the state and the practice of outsourcing to data annotation institutions. It asks the following two research questions: what is the role of the state in the US and China's data annotation industry and how is the process of data annotation organised by tech companies within each country?

We argue that while the US state plays a protectionist role in securing AI infrastructure such as chip manufacturing and cloud computing facilities, it does not conceive of data annotation as a national strategic priority. The Chinese state, on the other hand, seeks to exercise considerable control over the institutional conditions of data annotation and co-manages a number of data annotation institutions alongside Chinese AI companies. We also find that outsourcing data annotation work to third-party countries is widespread among American and European firms, but that Chinese language data annotation tasks of Chinese companies are more likely to be completed in-house or through a practice of 'inland-sourcing' to workers in 'third-tier' Chinese cities, concepts which we explain in greater detail below.

We argue this research has three main implications for our understanding of data annotation and the global development of AI. The first is that existing frameworks of data colonialism, digital colonialism and decolonial AI need to be updated to integrate the position of China as a global digital superpower with its own unique practices and organisational forms that differ significantly from the USA and Europe. China's leading position as the second largest investor in AI and a significant player in the global race for AI currently fits awkwardly within theoretical frameworks developed from the Latin American modernity/coloniality school that privilege the practices of European colonialism in the creation of new modalities of digital exploitation and imperialism today (Kwet, 2019; Mohamed et al., 2020; Muldoon and Wu, 2023). Instead, we argue that the USA and China now compete in a largely bipolar world of AI development in which the two major superpowers utilise different models of economic investment and political oversight. While one can trace legacies of colonialism in the patterns of power that shape technology development in the USA and Europe, China has its own political history and geopolitics that influence its model of AI development that requires further analysis. The USA and China understand themselves as in competition and must mobilise other middle powers such as Saudi Arabia, the UAE, South Korea and the UK in their struggle for hegemony.

Second, our analysis calls for renewed attention to geopolitical considerations and the role of the state in the study of data annotation, with previous studies focussing more on employment relations and working conditions of data annotators. We show that state policies play a significant role in shaping the institutional conditions of data annotation, which determine the organisational forms that data annotation adopts in different national contexts.

Third, the article points to a major difference between how datasets and data annotation are considered by the two leading digital superpowers, emphasising that China's belief in trailing the USA in AI development incentivises it to prioritise data as an area in which it holds a strategic advantage as the largest producer of global data. While the USA prioritises chip manufacturing and AI data centres, the Chinese government invests heavily in and seeks to co-manage data production institutions as a means of catching up to the USA in the AI race.

Artificial intelligence and data annotation

The data annotation industry has grown significantly with the explosion of interest in AI and the need for more high-quality data (Miceli et al., 2022b). Data annotation involves significant amounts of human labour by data workers who curate, annotate and evaluate datasets that are used to train machine-learning algorithms. We define AI data work as 'the human labour required to support machine learning algorithms through the preparation and evaluation of datasets and model outputs that is often outsourced to low-paid and marginalised workers' (Muldoon et al., 2024a). Due to the costly and labour intensive nature of data annotation, many AI companies choose to outsource these tasks to external providers (Tubaro and Casilli, 2019). Geographers have demonstrated that many of these AI companies are located in the Global North and outsource data work to the Global South due to weaker labour laws and the lower costs of labour (Gray and Suri, 2019).

Literature in the social sciences has focussed on the working conditions and employment relations of these workers, demonstrating that this outsourced work is often performed for low pay with insecure contracts and limited communication between workers and AI companies (Miceli and Posada, 2022; Muldoon et al., 2023). Labour-intensive data annotation has been found to be a 'structural feature' of the production of AI, which is set to expand with the further growth of the sector (Casilli, 2025; Tubaro and Casilli, 2019). A large literature in computer science and human-computer interactions has also argued that the interventions of humans in the data production process can produce certain biases due to the subjective positions of workers and the interpretive tasks they must perform (Diaz et al., 2022; Miceli et al., 2020; Muller et al., 2019; Passi and Jackson, 2018; Paullada et al., 2021). Miceli and Posada (2022) argue through the concept of a 'data-production dispositif' that data annotation is embedded in certain hierarchical conditions of knowledge production that shape the datasets and AI systems produced using human annotation.

Data annotation is most often performed through two types of data work institutions: digital platforms and BPOs (Muldoon et al., 2024b; Posada, 2021). Some of the earliest forms of online gig work were facilitated by generalist microwork platforms such as Amazon Mechanical Turk in which the platform would operate online marketplaces for digital tasks, allowing requesters to post various jobs to a distributed workforce (Berg et al., 2018; Howcroft and Bergvall-Kåreborn, 2019). These platforms enable companies to access a large workforce at a relatively low cost, suitable for large-scale data categorisation and annotation tasks. However, security and privacy concerns arise due to the lack of oversight on data accessed by thousands of distributed workers. Quality control is also a significant issue with digital platforms due to the cyclical nature of data annotation and the lack of direct communication between the AI company and workers (Le Ludec et al., 2023; Muldoon et al., 2024a).

Although the literature on digital platforms is much larger, AI data work is now more likely to be performed at BPOs where workers have employment contracts and work at a particular centre for months or years at a time (Le Ludec et al., 2023; Posada, 2021). Miceli and Posada (2022) define a BPO as 'a form of outsourcing that involves contracting a third-party service provider to carry out specific parts of a company's operations, in the case of our investigation, data-related tasks'. The BPO sector first emerged in the USA and Europe in response to competitive pressures to restructure business processes for increased efficiency by outsourcing tasks to lower-wage areas within national economies (Davis, 2009). The BPO industry grew rapidly in the 1990s and 2000s due to the global spread of the Internet and ICT services, reducing communication costs and enabling international partnerships. Studies have demonstrated that BPOs provide several distinct advantages over crowdsourcing platforms. First, their management structure allows closer supervision of workers, resulting in higher quality outputs (Le Ludec et al., 2023). Detailed task instructions can be communicated through the BPO hierarchy, ensuring consistent and accurate work. BPOs can also be more specialised than digital platforms and can focus on specific types of data services, such as those needed for autonomous vehicles or computer vision. They can also offer end-to-end services, managing multiple stages of a client's data needs, enhancing output quality, and better managing data at each stage (Muldoon et al., 2024a). They are typically more expensive than digital platforms because they offer higher service quality, direct communication lines between clients, management, and workers, and a higher level of information security.

While China's data annotation operations exhibit distinct characteristics, empirical studies are scarce. This scarcity is partly due to the challenges researchers faced in gaining access and collecting data during the strict pandemic control period. Nevertheless, Chinese scholars such as Wu, Xia and Hu have conducted fieldwork during the pandemic years and have begun publishing a series of studies that explore data annotation operations beyond the theoretical frameworks of digital platforms and BPOs (Hu, 2023; Wu and Xia, 2023; Xia and Wu, 2025). Wu and Xia's work identifies various 'data labelling bases' (DLBs) in China, demonstrating these bases as collaborative efforts between tech giants and local actors in third- and fourth-tier cities,[1] including local governments, NGOs, and vocational schools. In this model, tech firms establish the bases, while local actors secure subsidies, maintain the labour pool, and manage workers (Wu and Xia, 2023). Drawing on the case of China's medical image annotation, Hu (2023) illustrates that some of these annotation tasks require professionals with relevant medical expertise, challenging the prevailing Western assumption that annotation is low-skilled work requiring minimal training.

Building on the literature concerning China's AI data industry, we address several gaps. First, we argue that existing studies have largely neglected the state's role in shaping this sector, an important oversight given the Chinese state's active involvement in guiding technological development (Lei, 2023). Our study also incorporates a comparative analysis of China and the USA, situating China within the global data annotation system and challenging the traditional core-periphery dichotomy (USA/Western Europe vs. Global South). This comparative analysis also reveals distinct organisational forms in China such as the DLB model instead of digital labour platforms and BPOs.

As shown above, the existing literature has a sociological focus on employment relations and working conditions and focuses on fieldwork conducted at US and European-connected digital labour platforms and BPOs. There is less focus on the larger political economy of data annotation as an industry or the industry's relationship

Table 1. Overview of field sites and interviews.

Site (pseudonym)	Type	Fieldtrip period (days)	Interviews/focus group (FG)
Factory A (Guizhou, China)	Data labelling factory (owned by a third-party provider, performing work for multiple companies)	July–August, December 2019 (13 days)	25 interviews, 3 FG
Base B (Guizhou, China)	Data labelling base (owned and operated by one tech company, performing work exclusively for them)	November 2019; November 2020 (38 days)	36 interviews, 3 FG
Company C (Dongbei, China)	Brokerage BPO (smaller operations often funded by venture capital and focussing on deploying new technology)	January 2020 (11 days)	25 interviews
Base D (Shanghai, China)	Data labelling base	December 2020 (16 days)	13 interviews
Base E (Shaanxi, China)	Data labelling base	January 2021 (9 days)	19 interviews, 1 FG
Base F (Guangdong, China)	Data labelling base	July 2021 (8 days)	7 interviews, 2 FG
Base G (Shanxi, China)	Data labelling factory	September 2020 (12 days)	14 interviews
Company H (Hangzhou, China)	Brokerage BPO	August–September 2020 (30 days)	11 interviews
Company I (Hangzhou, China)	Brokerage BPO	October–December 2023 (60 days)	11 interviews
Base J (Shanghai, China)	Data labelling base	October–December 2023 (70 days)	12 interviews
Factory K (Chongqing, China)	Data labelling factory	December 2023–January 2024 (10 days)	6 interviews

with other actors such as the state. When legislation is considered, it is primarily concerning legislation that would protect workers' rights. There are also geographic limitations with less focus on the Chinese data annotation ecosystem. Building on this existing literature, we extend this scholarship to offer a broader analysis of the data annotation ecosystem and compare what we know about existing data annotation practices with new empirical research on Chinese data annotation.

Methodology

We conducted a five-year multi-site ethnographic study (July 2019 to January 2024) across China to explore data annotation practices, evolving industry structures, and shifting state policies (see Table 1). Access to field sites was facilitated by key informants, such as annotators at Site D, algorithm engineers at Site I, and project managers at Site J, enabling research stays ranging from 8 to 70 days. Observations focused on annotators' daily tasks, organisational practices (e.g., recruitment, managerial routines, training), and interactions (e.g., meetings, conferences) with local government and AI departments. Additionally, we collected policy documents and confidential training materials provided by participating firms.

To complement field observations, we conducted 192 semi-structured interviews with diverse participants across the AI data annotation ecosystem. These included:

- 35 managers or executives from AI firms, large third-party annotation facilities, BPOs, and DLBs;
- 23 engineers from algorithmic, operational, product, and data departments;
- 18 government officials and staff from national AI institutions, such as Data Bureaux, national labs, and digital parks;
- 14 technology journalists, NGO observers, and researchers with expertise in AI regulation;
- 102 data annotators, team leads, and supervisors at data annotation bases.

Annotators represented a diverse range of gender, age, migration status, and educational backgrounds. In China, annotators primarily fell into three groups: middle-aged women, workers with disabilities, and vocational school interns. Interviews lasted an average of 1–2 hours, with 16 conducted via video calls during pandemic-related restrictions. Recruitment combined purposive and snowball sampling, with participants identified either onsite during fieldwork or through networks facilitated by key informants and prior interviewees.

A thematic coding approach was employed to analyse over 200,000 words of fieldnotes and verbatim interview transcripts. Coding focused on patterns related to labour organisation, operational practices, and the interplay among state policies, annotation bases, and AI companies. To ensure reliability, two researchers independently coded a subset of transcripts and reconciled differences through discussion. Comparative analysis was conducted to contrast Chinese and US annotation ecosystems, highlighting both similarities and divergences.

This research adhered to ethical standards approved by institutional review boards. All participants provided

informed consent, were assured of confidentiality, and participated voluntarily. Participants were given detailed information about the study's objectives and methods, with all identifying information anonymised to protect their privacy.

In this article, we also conducted a literature review to determine the state-of-the-art research on American firms and their outsourcing practices in order to compare this with the Chinese fieldwork.

The role of the state: Securing AI infrastructure versus China's Manhattan project

China and the USA are the two leading tech superpowers engaged in a race to develop AI technology to boost their economic productivity and provide them with a military and strategic edge in geopolitics. Yet the role the state plays in the development of AI and in the data annotation industry is markedly different in each context. While the USA focuses on securing its own AI infrastructure such as access to cutting-edge AI chips and computing power from large data centres, the Chinese state plays a more interventionist role directly investing in tech companies, co-managing AI data annotation facilities and exercising its comparative advantage in developing high quality data sources. In this section, we show that while the US state adopts a largely laissez-faire approach to how its companies organise data annotation, the Chinese state strictly regulates the AI sector, including data annotation facilities.

Since the end of the Second World War, successive US administrations have understood the US science and technology sector as of critical importance to the geopolitical and economic ambitions of the state (Leslie, 1993; O'Mara, 2019). This sense of the strategic value of technology has only heightened in recent years with the rise of AI and the broadly held view among US policymakers that it is a foundational technology that will lead to rapid advances across multiple domains. Since the 1980s, US tech policy has focussed on promoting the tech sector abroad by adopting minimal regulations and providing subsidies and incentives for new technology companies (Kak and West, 2024; Thierer, 2014). The Biden administration changed course in its approach to industrial policy, with less emphasis on neoliberal principles of free market competition and greater levels of state-led investment and tax credits and incentive schemes. However, while spending and investment in particular aspects of AI have increased, the administration has maintained a similar regulatory strategy focussed on promoting innovation and fostering the AI industry (Szczepański, 2024).

US industrial policy on AI has focussed less on data annotation and more on investing in AI infrastructure such as AI chip manufacturing, cloud-based compute resources and the development of foundation models. This includes the CHIPS and Science Act of 2022, the NAIRR Pilot enacted by the National Science Foundation and the CREATE AI Act, which aim to subsidise US-based chip manufacturing and support the creation of public compute facilities for the research and development of AI (Kak and West, 2024). In terms of geopolitical importance, AI chips have taken the spotlight, with Biden announcing this sector as 'ground zero of our tech competition with China', and introducing restrictions on the sale of chips, equipment and other essential materials to Chinese manufacturers (White House, 2022a).

sWhile data and jobs have been important peripheral themes, data annotation has not been understood as key to the US AI policy. Under both the Trump and the Biden administrations, ensuring companies had access to 'AI ready' datasets has been a policy goal with efforts towards creating benchmarks for high quality data. This includes the benchmarks published by the Trump White House Office of Science and Technology Policy (OSTP) Subcommittee on Open Science and the NAIRR Task Force's plan for 'analysis ready' datasets (Long and Romanoff, 2023). Public investment in AI has also been directly linked to the creation of high quality jobs, such as with the Davis-Bacon prevailing wages provision of the CHIPS Act, requiring union jobs and minimum wages for any facilities that receive funding (White House, 2022b). However, these provisions have not been extended to any workers in the data annotation industry either in the USA or those based in other countries.

Although AI is viewed as a strategic national asset and certain aspects of its supply chain have received careful policy consideration, data annotation is not considered a bottleneck and is of less strategic importance than semiconductors and cloud computing. As a result, the outsourcing of data annotation services is left up to tech companies without additional regulations such as those that exist in other jurisdictions like the German Supply Chain Act or the European Corporate Sustainable Due Diligence Directive. These laws impose obligations on lead firms to ensure a set of minimum standards are maintained throughout their supply chain and to perform risk assessments to ensure compliance (European Commission, 2024). When it comes to the data annotation industry, the US state has largely adopted a laissez-faire approach, allowing companies to conduct their own annotation or outsource it. Some of this outsourced work has attracted negative media coverage that has focussed on the poor working conditions of workers in the Global South (Perrigo, 2023). However, the Biden administration's call for a 'worker-centric' approach to policymaking has not extended to data annotators outside of the USA who are not considered a strategic priority for US foreign policy. So, while the US state has adopted a protectionist policy towards its chips industry and has dedicated considerable public funds into AI infrastructure, it has remained on the sidelines in terms of data annotation

and has neither directly invested nor created a regulatory framework for the industry.

In China, the state plays a distinct role by heavily investing in AI as a national priority and tightly regulating its development. There are three key aspects to its strategy. First, AI is seen as a national project or primary importance that requires the full mobilisation of the state apparatus to support its development. Second, while the USA focuses on AI infrastructure such as chip manufacturing, cloud-based computing, and model development, China extends its attention to the development of the data industry and data-related work. Third, concerned about the potentially unregulated nature of the AI industry, the Chinese government plays a highly active role in overseeing all aspects of AI development, from determining who is permitted to develop models to enforcing stringent data policies and output censorship. The dynamics of these three core aspects of Chinese state policy create a tension which forces Chinese tech companies to embody the concept of 'dancing with shackles' [*dai liaokao wudao*] (Cai, 2023). According to this idea, tech companies are encouraged to play a leading role in a well–resourced field developing cutting-edge technology while at the same time being closely monitored and controlled by the Chinese government.

The Chinese government's emphasis on technological development, rooted in Deng Xiaoping's economic reforms that identified science and technology (S&T) as critical to productivity and growth, has intensified under subsequent leaders, particularly Jiang, Hu, and now Xi Jinping (Huang, 2008; Lei, 2023: 69–80). Building on the foundations of his predecessors, Xi has elevated the importance of S&T to unprecedented levels. He holds a steadfast belief that S&T will address many of China's challenges, including local governments' struggles with land- and labour-intensive industries, economic slowdowns, and domestic criticism. Under Xi's directive, the state actively invests in and regulates every facet of technological development, particularly AI.

The first key aspect of China's strategy is that it acknowledges that only by mobilising the entire national apparatus can it hope to rival the USA in AI development. In our fieldwork, we spoke with an experienced technology journalist who highlighted that AI development has now been elevated to a national strategic project, directly overseen by the State Council. After mentioning the State Council's involvement, he paused and asked us, 'Do you understand the implications of AI development being directly coordinated by the State Council?' When we admitted our uncertainty, he explained that this move signifies an unprecedented mobilisation of national resources: The State Council can control and coordinate efforts across all key departments, including the National Development and Reform Commission, the Ministry of Science and Technology (MOST), the Ministry of Industry and Information Technology (MIIT), and the Cyberspace Administration of China (CAC), to drive AI development. He concluded, 'This is China's Manhattan Project'.

Secondly, while the USA primarily focuses on critical AI infrastructure, the Chinese government places significant emphasis on the pivotal role of data and data annotation. Large, high-quality datasets are crucial for enhancing accuracy and generalisability in machine learning. OpenAI's recent struggle to secure sufficient training data for GPT-5 underscores the critical challenge posed by limited data (Alitech, 2023). Recognising its advantage in accessing vast datasets and a robust annotation workforce, China has strategically leveraged these resources to gain a decisive competitive edge in the global AI race. A recent National Bureau of Statistics report emphasised data as a new form of 'productivity' in China, with the country producing 8.1 ZB of data in 2022 – a 22.7% increase year-on-year, accounting for 10.5% of global production (National Bureau of Statistics of China, 2023). This is also confirmed by an algorithm engineer: 'Innovation in China? There's hardly any compared to the US. We mostly adopt open-source code from US companies. But our advantage is in data. By feeding immense volumes into our models, we can still achieve significant optimisation'.

The government has prioritised a shift towards accumulating 'high-quality' data – particularly annotated data – thereby fostering the development of the annotation industry. Additionally, the government views data annotation as a strategic response to rising unemployment in a post-industrial economy, where the manufacturing and service sectors no longer absorb enough of the workforce. Moreover, the spatial flexibility inherent in annotation work allows it to be channelled to inland provinces such as Shanxi, Shaanxi, Gansu, Xinjiang, and Guizhou – regions previously excluded from manufacturing export development due to geographic and transportation constraints. Since 2017, the government has integrated the annotation industry into its broader digital economy strategy, directing more tasks to these inland provinces.

To align its AI ambitions with national goals, the Chinese government has not only invested in the development of AI technologies but has also implemented a robust regulatory framework to ensure strict oversight. The Chinese government also plays a pivotal role in regulating all aspects of AI development, from directing and monitoring AI model creation to governing data, data annotation, and the development of state-operated generative AI for surveillance and governance purposes. This includes the *Interim Measures for the Management of Generative Artificial Intelligence Services* (effective 15 August 2023), which grants the central government control over which companies are licensed to provide generative AI services. In September 2023, the first batch of eight companies, including Baidu, ByteDance, and SenseTime, received approval, gaining a first-mover advantage. In pursuit of a competitive edge, many tech firms have been eager to comply

with regulations and obtain licences promptly; as a result, 11 additional AI large model products were approved on 4 November 2023. By March 2024, 117 generative AI models in China had completed registration. However, regulation also operates through restrictions – by April 2024, only 140 out of 305 registered models had been approved (Cyberspace Administration of China, 2025). Furthermore, the government maintains control through constant monitoring of registered AI models, imposing penalties, and even suspending models that fail to meet filing obligations.

The *Interim Measures* address three key aspects of data regulation, which significantly shape China's data annotation industry. First, they require AI service providers to conduct self-assessments for legal compliance and political sensitivity during both input and output stages. A research participant from a leading Chinese AI company revealed that, to meet these regulations, her firm has implemented robust safety mechanisms. One of the main strategies includes hiring additional content moderators to filter politically and legally sensitive material during both the prompt input and machine-generated output stages.

Second, the *Interim Measures*, alongside the *Data Export Security Assessment Measures and Application Guidelines*, provide detailed regulations on data acquisition and processing, with particularly strict controls on data export. The government is highly sensitive to the transfer of domestic data abroad and enforces strict measures on domestic data service companies. Key informants from major AI companies noted that these strict regulations have created 'double walls' – preventing international data from entering China and domestic data from being exported. This has made data outsourcing extremely difficult, leading to a system of 'inland-sourcing', discussed further in the next section.

Finally, the *Interim Measures* set explicit standards for data annotation. According to the regulation, AI model companies and service providers must establish clear, specific, and actionable annotation guidelines, provide comprehensive training to annotators, and ensure the traceability, safety, and accuracy of annotated data. These stringent requirements have led many companies to strengthen their ties with data annotation bases, resulting in a trend of company-owned facilities and a streamlined data supply chain to improve traceability.

Outsourcing and 'inland-sourcing'

Lead tech firms in the USA and Europe often engage in outsourcing low-paid and low-skill data annotation work to suppliers based in the Global South. In contrast to the model training performed by machine learning engineers within the AI labs of the largest tech firms, the labour-intensive data annotation required to develop AI systems is often outsourced to suppliers where labour costs are lower and labour laws are more flexible (Miceli and Posada, 2022; Muldoon et al., 2024b; Tubaro and Casilli, 2019). Unlike their counterparts in the USA and Europe, Chinese firms seldom outsource annotation tasks to the Global South. Instead, the predominant organisational model for data annotation is the DLB, a partnership between a leading tech firm and local government or non-government actor. In this model, workers are employed by the local organisation (governmental or commercial) but work full-time exclusively on projects for the leading tech company. In a process we call 'inland-sourcing', major Chinese AI firms on the East coast – such as Baidu and Bytedance in Beijing, Alibaba in Hangzhou, and Tencent in Shenzhen – channel their annotation tasks to DLBs in inland provinces like Shanxi, Shaanxi, Gansu, Xinjiang, Guizhou, Chongqing, and Henan.

Several factors drive the inland-sourcing model in China. Primarily, China is highly protective of data, given concerns over political sensitivity and confidentiality, leading to the creation of a 'double wall' – blocking international data from entering the country and preventing domestic data from being exported. Additionally, data annotation serves as a strategic tool to boost China's digital economy, create employment, and bridge the development gap between the eastern coast and inland regions. Both central and local governments, eager to benefit from digital economic growth, offer favourable policies and subsidies to encourage tech firms to send annotation work inland, integrating the data industry into broader digital economy initiatives.

Data labelling bases

Our fieldwork identified three primary approaches to 'inland-sourcing': sending work to (1) tech companies' own DLBs, (2) large third-party annotation facilities or 'data labelling factories', and (3) brokerage BPOs. In the first approach, tech companies establish annotation bases across multiple inland provinces, integrating them into their digital development programmes, while a central data-coordinating department operates from a headquarters on the eastern coast. By the end of our first ethnographic fieldwork phase in 2021, B-Tech – recognised as one of the top four tech companies – had set up six bases in Guizhou, Shanxi, and Shaanxi provinces, with plans to expand by establishing an additional 17 bases in other inland regions. Annotation tasks were distributed flexibly across these bases, allowing the company to optimise task allocation and scheduling via a centralised platform. This system ensured that tasks were matched to the most suitable base, but also fostered competition among the bases, similar to the competition that occurs between BPOs competing for tasks from Western tech companies.

In our empirical cases, Sites B, D, E, F, and J all can be categorised as DLBs. Several key features define these annotation bases. First, these DLBs are exclusively owned and operated by a specific tech corporation, serving only that

firm. For instance, in our empirical cases, Base B operated exclusively for B-Tech (pseudonym), while Bases D, E, and F were dedicated solely to M-Tech (pseudonym), and Base J served X-Tech (pseudonym). Each base maintains a stable workforce, typically ranging from 50 to 100 workers, supported by a potential labour pool within the surrounding community, such as villages, digital economic zones, or poverty alleviation areas. Workers are directly hired by the tech firms, with contracts offering a monthly salary of 1500 yuan (approximately US$209) and bonuses based on a piece-rate system. Tech firms and local governments attempt to maintain stable and long-term relationships by establishing bases with such conditions as at least three years of rent exemption for the tech companies and the government covering the operating costs of the bases, followed by another five years of rent subsidies and reduced operating costs.

Even with local government subsidies, operating a DLB with a stable labour force is often more costly than outsourcing to BPOs. This raises the question of why many major AI companies in China adopt this model. The primary concern is data confidentiality. While Chinese tech firms may lag behind AI labs like OpenAI in algorithm innovation, they hold a significant advantage in data pipelines. Protecting this data is crucial to maintaining their competitive edge. An engineer from a Chinese AI company explained,

> Big Tech, of course, doesn't want to put their data-labelling tasks on an open platform. After all, the data a company labels reveals what they plan to develop next. So, especially for the big tech firms, they're really concerned about keeping their data confidential and they prioritise this over everything else….

Consequently, many large tech firms prefer to establish their own DLBs in less developed inland villages, where high-trust kinship networks and limited external connections help ensure the protection of sensitive data.

Taking Base B as an example, this DLB was established by B-Tech in 2018 within a poverty-alleviation relocation community in Guizhou. The DLB resulted from collaboration between local government and tech capital: the local government offered a three-year rent exemption and subsidies (e.g., for training and operations), while B-Tech committed to creating local employment opportunities. At the government's request, priority was given to marginalised women, including those with limited education, middle-aged individuals, and divorcees. Introducing the 'significance' of Base B, a team leader proudly stated:

> We're definitely the direct troop of B-Tech. For certain data, like level-3 confidential data, B-Tech would never let other contractors or suppliers handle it – it has to be us. For example, we recently worked on a project for the United Nations, one of those international AI projects. B-Tech took on the project and passed it directly to us…The project focused on identifying refugee shelters in a camp. For tasks at this level, B-Tech only trusts its own team.

We later learned that Level-3 data is classified as the high level of confidentiality by B-Tech. To help safeguard this data, the local government funded a secure 'confidentiality room' equipped with advanced surveillance systems within Base B. This facility enables workers to handle L3 data while allowing B-tech headquarter to monitor suspicious behaviour.

In this model, the ties between tech firms and their annotation bases are much stronger than those found in platform or BPO models in Europe and the USA. Unlike BPOs, which serve multiple companies simultaneously, the exclusive nature of DLBs allows them to focus entirely on collaborating with one particular tech company. This dedicated relationship fosters frequent and timely communication, ultimately ensuring superior data quality. One tech company – company M – that operates its own DLBs reported data accuracy rates of 97% to 98%, significantly exceeding the accuracy typically achieved by most BPOs it contracted.

Data labelling factories

In addition to company-owned DLBs, another type of organisation is emerging within China's data annotation industry of much larger facilities or 'data labelling factories' (DLFs) owned by third-party providers, exemplified by our empirical cases, Factory A in Guizhou and Factory K in Chongqing. These factories operate on a much larger scale, typically employing between 500 and 1000 workers, who sign contracts with the third-party companies rather than specific tech firms. The workforce is notably young – Factory A employs exclusively vocational school students through on-the-job internships (*ding gang shi xi*), while Factory K comprises mostly of recent graduates with slightly higher education qualifications. Unlike company-owned DLBs that serve only their parent firm, these third-party factories handle tasks for multiple tech companies simultaneously. For example, Factory K simultaneously handled annotation tasks for Tencent, ByteDance (parent company of TikTok), Didi (China's leading ride-hailing tech company), Huawei. Typically located in inland regions, they are integrated into provincial digital economic development programmes, which helps reduce startup and infrastructure costs. For example, Factory K we visited in our fieldwork employs over 500 annotators and would have needed to invest heavily in renting or building workspace. However, through collaborative relations with Chongqing, the local government provided four floors of a skyscraper constructed as part of their special digital development zone. According to one of Factory K's managers, with the free use of these four floors, they could easily expand their workforce to 1000 in the coming years.

The advantage of these third-party providers lies not in handling confidential data but in managing more

sophisticated and specialised data. The scale of these factories, combined with advanced recruitment, more stable job ladders and training processes, ensures this capability. First, leveraging their extensive labour pool, these factories employ targeted recruitment strategies to identify the most suitable annotators for specific tasks. For instance, Manager Zhang, who coordinated all projects at Factory K, explained that the company tailors its recruitment criteria and interview questions to match the requirements of each annotation project:

For different projects, we hire based on specific needs. For math-related tasks, we ask

> HR to filter candidates with their national exam math grades, and the project team creates math-focused interview questions…If we accepted clients' language tasks… like writing conversational text to mimic animated characters, we prefer candidates with language or literature majors.

These recruitment strategies undoubtedly establish a foundation for assembling a workforce capable of managing increasingly sophisticated and specialised tasks, similar to DLBs. However, retaining such workers – particularly those with higher educational qualifications, such as bachelor's degrees – in a job characterised by relatively low pay and limited prestige presents significant challenges. Consequently, the creation of a stable career progression pathway becomes crucial, as highlighted by Manager Xia from Factory F:

> our company has three-tier leadership structure: team leader at the bottom, then project supervisor, and project manager at the top. If you do a good job, you get promoted …for students from universities with bachelor's degrees, it can be tricky because their salary expectations are high. To attract them, we recruit through a management trainee program (*guan pei*) and explain that if they perform well, they'll be fast-tracked to managerial roles.

Also, extended training sessions on annotation tasks, led by more experienced and skilled workers, are critical to a company's capacity to manage sophisticated and specialised data. For complex data, training can often take several weeks. For instance, Supervisor Zhang, who is currently in charge of training at Factory F, explained: 'we needed to spend half a month training people in '3D points cloud annotation'. And it takes at least two to three months for them to really get the hang of it'. The challenge is that workers' productivity tends to be very low during the training period, often falling short of earning even a minimum salary. These large factories accommodate the extended training period through internal coordination and payroll flexibility, laying a solid foundation for their workers' ability to undertake sophisticated annotation work. The ability to handle highly specialised and complex tasks gives these factories substantial negotiating power with their client tech firms. For instance, in dealing with LLM annotation – far more complicated and experimental than traditional types – each batch of data often requires extensive rework. If paid on a piece-rate basis the need for constant revisions would greatly reduce the factory's earnings. To address this, Factory K we visited successfully negotiated with its client to ensure workers were paid monthly salaries, mitigating the risk of excessive rework without additional compensation.

Brokerage BPOs

The final type of data annotation closely resembles the BPO facilities of Europe and the USA, albeit on a much smaller scale, typically employing around 20 people. These brokerage BPOs are defined by two key characteristics. First, they are often venture capital-driven. All BPOs we studied had completed their Pre-A financing rounds by 2023. Second, these BPOs focus heavily on the improvement and optimisation of annotation technology, a priority shaped by venture capital demands. This emphasis on cutting-edge technology sets them apart from the two other types of facilities previously discussed. At our field site, Company I, located in Hangzhou with around 20 employees and recently completing its Pre-A funding, half the workforce focused on engineering tasks, particularly automating the annotation platform. By the end of our fieldwork, Company I had successfully integrated its automation algorithm into the 'pre-annotation' process. This innovation automates initial annotation tasks, significantly reducing outsourcing labour. Manager Huang from Company I elaborated:

> Now, when the images are uploaded to the platform, the system immediately deploys its automation algorithm… critical information on the road, such as pedestrians and barricades, is directly annotated by the machine… the only task that requires human intervention is correcting some corner cases.

China's BPOs, much like their US and European counterparts, have played a significant role in shaping the AI supply chain, often contributing to its opacity. Two common strategies employed by BPOs to influence the supply chain include concealing supplier information and extending supply chains. These BPOs have obtained approval from major tech companies to ensure a 'blind-eye' approach to supply chain management in which tech companies only deal with the BPO and allow them to subcontract the work to other parties without interrogation by the leading tech company. BPOs often take this work and then subcontract it out to companies and regions with even lower labour costs in order to make more profit.

By relinquishing control over further subcontracting, tech companies have allowed BPOs to extend the AI supply chain even further into lower-tier cities to minimise labour

costs. As one manager at Company H, a typical brokage BPO located in Hangzhou, explained:

> In third-, fourth-, and even fifth-tier cities, there is a strong desire to catch up with AI developments. However, it is challenging [for them] to engage in advanced AI algorithm development. In this context, the data annotation business could be very appealing if it sinks into these lower-tier cities.

Due to this supply chain extension, another critical player in this chain is the 'linkage broker' (*chuan chuan*). One manager at an AI firm described the role of linkage brokers as follows:

> In the fourth-tier and fifth-tier cities of Henan and Hebei, there are many linkage brokers. They can always find cheaper locations. They register as liaison points and tirelessly negotiate deals from their upstream linkage brokers, then bring them back to their downstream suppliers and redistribute them.

These linkage brokers, strategically positioned across inland provinces like Henan and Hebei, ensure that jobs can often be further outsourced to different providers. This process pressures fourth- and fifth-tier cities to compete by increasing government support, lowering labour costs, and relaxing labour regulations to retain annotation work.

The extended supply chain has resulted in a decline in wages and a deterioration of working conditions. As one manager explained:

> Initially, when highly educated individuals handled the annotation work, it was quite expensive… I remember the rates could go up to over ten thousand yuan. Back in 2015 and 2016, the wages for annotators were extraordinarily high… which motivated us to find ways to reduce labour costs during those early days.

As automation algorithms began to handle more complex annotation scenarios, the demand for annotators with high comprehension and quick learning abilities significantly decreased within BPO labour pools, leading to a substantial decline in wages. During our visit to one BPO subcontractor, we found that wages had dropped to between 1500 and 3000 yuan per month.[2] Across the sector at all types of data annotation facilities in China, we observed a decrease in wages of data annotators over the past 5-year period. In sum, the organisational forms of DLBs, DLFs and Brokerage BPOs provides a different context for the provision of data annotation services in the Chinese case that differs significantly from the US and European outsourcing models.

Conclusion

The comparative analysis of the AI data annotation ecosystems in the USA and China reveals significant structural and organisational differences including the role each state plays in the AI data annotation industry and how firms organise the process of data annotation. As the two leading digital superpowers, the USA and China adopt divergent strategies when it comes to the cultivation of their AI data annotation ecosystems. American AI firms follow the path of previous generations of US companies in outsourcing core parts of the development process to countries in the Global South, particularly when it comes to labour-intensive data annotation. The Chinese state exhibits a greater attentiveness to its data ecosystem than in the USA where focus is on chip manufacturing and compute resources. Instead of allowing for data to be sent abroad, the Chinese state seeks to directly cultivate a data ecosystem inside its borders where data annotation and the management of datasets can be undertaken by companies with strong ties to local and national government. We have seen that both heightened security considerations and a sense of holding a relative advantage in data has led the Chinese government to invest more heavily in producing data than in the USA.

The emerging position of China as a leading technology superpower opens questions of how its practices compare to what scholars have taken to be new forms of digital and data colonialism. Building on the work of the Latin American modernity/coloniality school, these scholars have analysed how even with advanced technology, there is an international division of digital labour between workers in core and peripheral countries in which different types of labour with varying levels of economic and social capital are assigned to different groups of workers (Quijano, 2000; Muldoon and Wu, 2023). One limitation with this framework is that it theorises these economic orders as an outgrowth and extension of European colonialism, which creates a dichotomy between the Global North (USA, Western Europe) and the Global South, which makes it difficult to integrate China's emerging status as a global technology superpower.

We find that the question 'is China a new digital colonial power?' to be inappropriate because the analytical framework of colonialism and neo-colonialism relates to a particular historical and geographic conjuncture that fails to capture the historical experience of China's rise and its current position (Lee, 2018). Building on the scholarship of Ching Kwan Lee, we emphasise that China's role in this phase of digital globalisation differs markedly from the earlier 'Global China' regime, characterised by an aggressive strategy of direct investment in countries in the Global South over the past two decades. The current phase of AI development is more inward looking and is marked by attempts to develop home grown alternatives to US technology.

In the case of data annotation and AI development more generally, China's inward stance is notable within the current wave of digital globalisation. In the USA, the presence of a liberal government, high wages and limited regulation incentivises firms to reduce costs by outsourcing data

annotation work to BPOs in the Global South. Without the intervention of other political priorities, US firms focus on profit maximisation, which is achieved through locating the cheapest and most efficient form of data annotation from a range of global suppliers. In China, the strategic outlook of firms is complicated by an interventionist and regulatory state, which actively seeks to direct the development of AI, at times prioritising security concerns or the competitive pursuit of the AI race over profitability.

Indeed, when perceiving China's position in current wave of global AI development, the pivotal and distinct role of the interventionist state must be taken into account. Unlike the exploitative structures of Western data colonialism, China's approach integrates developmental, sovereign, stability and ideological goals. On one level, the Chinese state seeks to advance its developmental goals and dominate the global AI landscape by leveraging the country's vast reserves of data. Inland sourcing enables the establishment of a self-contained, domestically driven data ecosystem that safeguards national data sovereignty and prevents Western extraction. This ambition is evident in the stringent regulatory oversight of the data chain and the promotion of inland-sourcing. At the same time, controlling data annotation aligns with the state's overriding priority: maintaining social and political stability. In China's one-party system, stability forms the foundation of regime survival and is consistently prioritised above economic profitability. This imperative drives direct state intervention in data work, particularly as it involves handling politically sensitive content and aligning machine outputs with national ideology. Inland sourcing thus emerges as a deliberate strategy to enhance state oversight, counteract Western ideological influence, and reinforce ideological control.

China faces certain limitations in outsourcing in that unlike English and French speaking countries, there are not as many large Chinese-speaking low-income countries to receive outsourcing work. However, language abilities are not required for many computer vision data annotation tasks and we consider that the security and political objectives are perhaps the more relevant factor in why this work has not been outsourced as much as with US firms. Such balanced considerations are reinforced by China's unique organisation of annotation work. For instance, our findings reveal middle-aged women are prioritised for recruitment in DLBs through partnerships between local governments and tech companies. Similarly, student interns form the primary labour source for data labelling factories via vocational schools' on-the-job training programmes. These organisational patterns actively involve local actors, such as governments and vocational schools, to integrate more disadvantaged labour groups into the industry. The presence of government actors in this space brings added security to these deals, which has enabled longer-term contracts between parties, increased spending on establishing DLBs and the further development of the sector. The geopolitical strategy of each country has led to different policy positions in the development of AI and alternative ways of conceptualising the state's relationship to its data ecosystem.

The implications of these findings are significant for understanding the global political economy of AI development. Firstly, the study challenges the prevailing theories of data colonialism and digital colonialism, which often focus on the exploitative practices of US and European firms that outsource data annotation tasks to the Global South. These theories, grounded in Latin American modernity/coloniality frameworks, have tended to overlook China's position as a global digital superpower. China's practices and organisational forms differ markedly from those of the West, suggesting that theoretical models of digital exploitation need to be updated to account for China's unique role in the global AI industry.

Moreover, the role of the state in shaping the institutional conditions of data annotation is a crucial factor that has been underexplored in previous studies. The USA and China represent two distinct models of state involvement in AI data annotation. The USA, with its laissez-faire approach, allows private companies to determine the organisational form of data annotation, often outsourcing it to regions with cheap labour. In contrast, China's more centralised and coordinated approach involves significant state intervention, with the government working alongside private firms to ensure the development of a robust domestic data ecosystem. This contrast illustrates the broader geopolitical dimensions of AI development, where state policies and priorities shape not only the structure of the AI industry but also the conditions under which data annotation labour is performed.

China's decision to prioritise data production as a strategic advantage in the global AI race also highlights a fundamental difference in how datasets and data annotation are valued by the two leading AI powers. The US government has focused its efforts on protecting AI infrastructure, believing this to be the critical component of its current lead in AI development. China, on the other hand, sees data as an important resource that gives it a unique advantage, particularly as the country produces vast amounts of data that can be used to train AI models. This strategic prioritisation of data aligns with China's broader ambitions to catch up to and potentially surpass the USA in AI development. Ultimately, the findings of this article suggest that the global AI race is not merely a competition between private companies, but also a contest between different state-driven models of AI development. The USS and China represent two distinct approaches, with the USA relying on private sector innovation and outsourcing labour, while China adopts a more state-driven, coordinated approach that emphasises domestic control over data and labour. These differences are likely to shape the future trajectory of AI development and will have implications for broader geopolitical questions of global security and economic development.

ORCID iDs

Tongyu Wu https://orcid.org/0000-0003-1184-4498
James Muldoon https://orcid.org/0000-0003-3307-1318
Bingqing Xia https://orcid.org/0000-0002-6020-9960

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Notes

1. The 'third tier' and 'fourth tier' ranking refers to an unofficial hierarchical classification of Chinese cities often used by investors to refer to the different levels of infrastructure, commercial enterprises, size, location and consumer behaviour of different Chinese cities.
2. According to the supervisor, wages are divided into two tiers. Intern workers receive a basic wage of 1,500 yuan per month, along with a 200 yuan attendance bonus and 300 yuan in other subsidies. Full-time workers receive a basic wage of 3,000 yuan, with project commissions capped at 1,000 yuan.

References

Alitech (2023) OpenAI runs out of training data for GPT-5: Running out of high-quality data and its impact. Available at: https://alitech.io/blog/openai-runs-out-of-training-data-for-gpt-5/ (accessed 25 January 2025).

Berg J, Furrer M, Harmon E, et al. (2018) Digital labour platforms and the future of work: Towards decent work in the online world. Report. Geneva: ILO. Available at: http://www.ilo.org/global/publications/books/WCMS_645337/lang–en/index.htm (accessed 31 July 2024).

Cai J (2023) AI Developers Must Learn to 'Dance in Shackles' as China Makes New Rules for Post-ChatGPT World. South China Morning Post, 24 April 2023. Available at: https://www.scmp.com/news/china/politics/article/3218144/ai-developers-must-learn-dance-shackles-china-makes-new-rules-post-chatgpt-world (accessed 27 January 2025).

Casilli A (2025) *Waiting for Robots: The Hired Hands of Automation*. Chicago: University of Chicago Press.

Casilli A and Tubaro P (2019) Artificial intelligence as the new frontier of capitalism: A socio-economic critique. In: Pichault F and Pélissier T (eds) *The Transformation of Work in the Digital Era*. London: Edward Elgar, pp. 63–78.

Cyberspace Administration of China (2025) Generative artificial intelligence service license released. cyberspace administration of China, April 2, 2024. Available at: https://www.cac.gov.cn/2024-04/02/c_1713729983803145.htm (accessed 27 January 2025).

Davis GF (2009) *Managed by the Markets: How Finance re-Shaped America*. Oxford: Oxford University Press.

Díaz M, Kivlichan I, Rosen R, et al. (2022) Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *Proceedings of the 2022 ACM Conference on Fairness*, Accountability, and Transparency. 2342–2351.

European Commission (2024) Corporate sustainability due diligence. Available at: https://commission.europa.eu/business-economy-euro/doing-business-eu/sustainability-due-diligence-responsible-business/corporate-sustainability-due-diligence_en (accessed 31 July 2024).

Gray M and Suri S (2019) *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. New York: Houghton Mifflin Harcourt.

Howcroft D and Bergvall-Kåreborn B (2019) A typology of crowdwork platforms. *Work, Employment and Society* 33(1): 21–38.

Hu W (2023) Machine learning in medical systems: Toward a sociological agenda. In: Schüßler U, Whittington R and Krücken G (eds) *The Oxford Handbook of the Sociology of Machine Learning*. Oxford: Oxford University Press, pp. 483–508.

Huang Y (2008) *Capitalism with Chinese Characteristics: Entrepreneurship and the State*. Cambridge: Cambridge University Press.

Kak A and West EM (2024) A modern industrial strategy for AI?: Interrogating the US approach. AI Now Institute. Available at: https://ainowinstitute.org/publication/a-modern-industrial-strategy-for-aiinterrogating-the-us-approach (accessed 31 July 2024).

Khanna S and Chandran R (2023) Gigs, scams, ghost work: India tech sector's dark side. *Reuters*, 26 March. Available at: https://www.reuters.com/world/india/feature-gigs-scams-ghost-work-india-tech-sectors-dark-side-2023-03-26/ (accessed 28 April 2025).

Knight W (2017) *China Plans to Use Artificial Intelligence to Gain Global Economic Dominance by 2030*. MIT Technology Review. Available at: https://www.technologyreview.com/2017/07/21/150379/china-plans-to-use-artificial-intelligence-to-gain-global-economic-dominance-by-2030/.

Kwet M (2019) Digital colonialism: US empire and the new imperialism in the global south. *Race Class* 60(4): 3–26.

Lee CK (2018) *The Specter of Global China: Politics, Labor, and Foreign Investment in Africa*. Chicago and London: University of Chicago Press.

Lei YW (2023) the gilded cage: technology, development, and state capitalism in China.

Le Ludec C, Cornet M and Casilli A (2023) The problem with annotation. Human labour and outsourcing between France and Madagascar. *Big Data & Society* 10(2): 20539517231188723.

Leslie SW (1993) *The Cold War and American Science: The Military-Industrial-Academic Complex at MIT and Stanford*. New York: Columbia University Press.

Long S and Romanoff T (2023) AI-ready open data. Bipartisan Policy Center. 17 February 2023. Available at: https://bipartisanpolicy.org/explainer/ai-ready-open-data/ (accessed 31 July 2024).

Maslej N, Fattorini L, Perrault R, et al. (2024) Artificial Intelligence Index Report 2024. Stanford University: Human-Centered Artificial Intelligence. Available at: https://aiindex.stanford.edu/report/ (accessed 28 April 2025).

Miceli M and Posada J (2022) The data-production dispositif. arXiv:2205.11963. arXiv. DOI: 10.48550/arXiv.2205.11963.

Miceli M, Posada J and Yang T (2022a) Studying up machine learning data: Why talk about bias when we mean power? In Proceedings of the ACM on Human-Computer Interaction 6, GROUP (2022), 1–14.

Miceli M, Schuessler M and Yang T (2020) Between subjectivity and imposition: Power dynamics in data annotation for computer vision. In *Proceedings of the ACM on Human-Computer Interaction 4*, CSCW2 (Oct. 2020), 1–25.

Miceli M, Yang T, Alvarado Garcia A, et al. (2022b) Documenting data production processes: A participatory approach for data work. In *Proceedings of the ACM on Human-Computer Interaction 6(CSCW2)*: 510:1-510:34. DOI: 10.1145/3555623.

Mohamed S, Png M-T and Isaac W (2020) Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology* 33(4): 659–684.

Muldoon J, Cant C, Graham M, et al. (2023) The poverty of ethical AI: Impact sourcing and AI supply chains. *AI & Society* (40): 529–543.

Muldoon J, Cant C, Wu BA, et al. (2024a) A typology of artificial intelligence data work. *Big Data & Society* 11(1): 1–13.

Muldoon J, Graham M and Cant C (2024b) *Feeding the Machine: The Hidden Human Labour Powering AI*. London: Canongate.

Muldoon J and Wu BA (2023) Artificial intelligence in the colonial matrix of power. *Philosophy & Technology* 36(4): 1–24.

Muller M, Lange I, Wang D, et al. (2019) How data science workers work with data: Discovery, capture, curation, design, creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). Association for Computing Machinery*, Glasgow, Scotland UK, 1–15. https://doi.org/10.1145/3290605.3300356.

Murgia M (2024) *Code Dependent: Living in the Shadow of AI*. London: Picador.

National Bureau of Statistics of China (2023) Data development and governance in the AI era. Available at: https://www.sohu.com/a/733822239_121106854 (accessed 31 July 2024).

O'Mara M (2019) *The Code: Silicon Valley and the Remaking of America*. New York: Penguin Press.

Passi S and Jackson SJ (2018) Trust in data science: Collaboration, translation, and accountability in corporate data science projects. In *Proceedings of ACM Human-Computer Interaction 2*, CSCW (Nov 2018), 1–28. https://doi.org/10. 1145/3274405 (accessed 31 July 2024).

Paullada A, Raji ID, Bender EM, et al. (2021) Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns* 2(11): 100336.

Perrigo B (2023) Exclusive: OpenAI USed Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic. Time Magazine. Available at: https://time.com/6247678/openai-chatgpt-kenya-workers/ (accessed 2 July 2024).

Posada J (2021) The coloniality of data work in Latin America. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 21 July 2021, 277–278. DOI: 10.1145/3461702.3462471.

Quijano A (2000) Coloniality of power and eurocentrism in Latin America. *International Sociology* 15(2): 215–232.

Schmidt FA (2019) Crowdproduktion von Trainingsdaten: Zur Rolle von Online-Arbeit beim Trainieren autonomer Fahrzeuge. Report, Hans-Böckler-Stiftung.

Shestakofsky B (2024) Cleaning up data work: Meaning, morality, and inequality in a tech startup. *Big Data & Society* 11(3): 1–14.

Szczepański M (2024) United States approach to artificial intelligence. European Parliament Research Services. January 2024. Available at: https://www.europarl.europa.eu/RegData/etudes/ATAG/2024/757605/EPRS_ATA(2024)757605_EN.pdf (accessed 31 July 2024).

Thierer A (2014) *Permissionless Innovation: The Continuing Case for Comprehensive Technological Freedom*. Arlington, VA: Mercatus Center at George Mason University.

Tubaro P and Casilli AA (2019) Micro-work, artificial intelligence and the automotive industry. *Journal of Industrial and Business Economics* 46(3): 333–345.

Tubaro P, Casilli AA and Coville M (2020) The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence. *Big Data & Society* 7(1): 2053951720919776.

Wang D, Prabhat S and Sambasivan N (2022) Whose AI dream? In search of the aspiration in data annotation. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems; 1–16.

White House (2022a) Remarks by President Biden in Meeting with CEOs and Labor Leaders on the Importance of Passing the CHIPS Act. 26 July 2022. Available at: https://www.whitehouse.gov/briefing-room/speeches-remarks/2022/07/26/remarks-by-president-biden-in-meeting-with-ceos-and-labor-leaders-on-the-importance-of-passing-the-chips-act.

White House (2022b) FACT SHEET: CHIPS and Science Act Will Lower Costs, Create Jobs, Strengthen Supply Chains, and Counter China. 9 August 2022, Available at: https://www.whitehouse.gov/briefing-room/statements-releases/2022/08/09/fact-sheet-chips-and-science-act-will-lower-costs-create-jobs-strengthen-supply-chains-and-counter-china (accessed 31 July 2024).

Wu T and Xia B (2023) Computing and manipulating: The confrontation between algorithmic and organizational control in the data labeling work. *Sociological Review of China* 11(6): 66–87.

Xia B and Wu T (2025) The space-time game: Workers with disabilities and mediating organizations in China's AI data labeling industry. *New Media & Society*. DOI: 10.1177/14614448251320114.