Check for updates

Original Research Article

BIG DATA & SOCIETY

Big Data & Society
January–March: 1–13
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20539517241232632
journals.sagepub.com/home/bds
Sage

A typology of artificial intelligence data work

James Muldoon[1], Callum Cant[1], Boxi Wu[2] and Mark Graham[2]

Abstract

This article provides a new typology for understanding human labour integrated into the production of artificial intelligence systems through data preparation and model evaluation. We call these forms of labour 'AI data work' and show how they are an important and necessary element of the artificial intelligence production process. We draw on fieldwork with an artificial intelligence data business process outsourcing centre specialising in computer vision data, alongside a decade of fieldwork with microwork platforms, business process outsourcing, and artificial intelligence companies to help dispel confusion around the multiple concepts and frames that encompass artificial intelligence data work including 'ghost work', 'microwork', 'crowdwork' and 'cloudwork'. We argue that these different frames of reference obscure important differences between how this labour is organised in different contexts. The article provides a conceptual division between the different types of artificial intelligence data work institutions and the different stages of what we call the artificial intelligence data pipeline. This article thus contributes to our understanding of how the practices of workers become a valuable commodity integrated into global artificial intelligence production networks.

Keywords

Artificial intelligence, microwork, crowdwork, data work, ghost work, business process outsourcing

Introduction

Artificial intelligence (AI) systems have recently attracted considerable attention in academia and the media and are quickly becoming embedded in several major platforms for work, communication and cultural production. These systems are designed with machine learning algorithms that leverage large amounts of data, energy and computational power (Bender et al., 2021; Crawford, 2021). In addition, AI systems require human labour – not only highly qualified machine learning engineers to design the algorithms, but also workers to assist with processing datasets (Gray and Suri, 2019; Tubaro and Casilli, 2019). Data used in machine learning algorithms must be collected, curated, annotated and evaluated by human workers. AI companies often don't have the resources or the expertise to ensure the quality and accuracy of this data work, at the scale that large-scale algorithms demand. Due to the labour-intensive nature and the cost of this process, most large technology companies outsource these tasks, either through business process outsourcing (BPO) or digital labour platforms, forming complex AI production networks (Miceli and Posada, 2022; Tubaro et al., 2020). These third-party data services provide flexible and affordable labour which is required to complete large-scale AI data work projects.

A focus on AI data work helps reframe techno-optimistic accounts of AI to centre the important role played by human labour in AI production networks, highlighting key aspects of the employment relations and working conditions that underpin them (Crawford, 2021; Dauvergne, 2022; Howcroft and Bergvall-Kåreborn, 2019). AI companies' data needs are growing considerably because 80% of hours spent on each AI project is estimated to consist of the collection, organisation and annotation of datasets (Cognilytica Research, 2019). The global data collection and labelling market size was estimated at $2.22 billion in 2022 and is expected to grow at a compound annual growth rate of 28.9% from 2023 to 2030 to reach $17.10 billion by 2030 (Grand View Research, 2022).

As this AI data work has expanded and a broader range of players have become involved in the industry, different terminology has proliferated including microwork,

[1]Essex Business School
[2]Oxford Internet Institute, University of Oxford, UK

Corresponding author:
James Muldoon, Political Science, University of Exeter, Stocker Rd, Exeter, UK.
Email: j.muldoon@exeter.ac.uk

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (https://us.sagepub.com/en-us/nam/open-access-at-sage).

crowdwork and ghostwork (Berg et al., 2018; De Stefano, 2016; Gray and Suri, 2019; Irani, 2015b). These terms can refer to similar, sometimes overlapping phenomena, but a degree of confusion currently exists within these debates as to which terminology should be employed to describe different arrangements of AI data work. There is a clear distinction, for example, between using geographically-distributed crowdsourced workers from a digital platform such as Amazon Mechanical Turk versus employing the services of a BPO company with employees based in an office with a traditional hierarchically organised managerial structure. Despite the growing interest in AI production networks and in the human labour that underpins machine learning algorithms, such distinctions are not always clear in the literature (Crawford, 2021; Miceli and Posada, 2022; Tubaro and Casilli, 2019). This article helps clarify these confusions by asking what are the different types of institutions that undertake data work and what set of interconnected processes is required to transform data into integrated datasets capable of being used in machine learning models?

In this article, we construct a new typology of AI data work which contains two important elements. First, we provide a table of *AI data work institutions* that distinguishes between six different ideal types of institutions through which AI data work is performed. This provides a heuristic device for determining the nature of AI data work institutions to help disentangle inconsistencies in the existing literature about what AI data work is and where it is performed. Second, we provide a conceptual schema for understanding the role of data work in the AI production process as a sequential system of steps in which data workers perform distinct roles along what we call an *AI data pipeline*. This adds to existing frameworks that have concentrated more on the functional roles of data workers in the AI production process (Tubaro et al., 2020). We adopt an organisational perspective to distinguish between how labour is performed under different employment structures and in different contexts as part of the production of AI.

In the construction of the first table on AI data work institutions, we undertake an examination of the existing literature and draw on a research project focused on AI training primarily for computer vision algorithms and applications. This project involved fieldwork conducted in 2023 with a BPO specialising in AI data work in Kenya and Uganda (Muldoon et al., 2023).[1] We also draw broadly on a wide body of fieldwork from our research team that has been collected since 2010. This includes three further projects: (1) fieldwork focused on the East African BPO sector, conducted between 2010 and 2014 (Graham, 2015); (2) research on remote work platforms, which included fieldwork in Vietnam, Malaysia, the Philippines, Kenya, South Africa, Ghana, Uganda and Nigeria (between 2014 and 2020) (Anwar and Graham, 2022; Graham et al., 2017); (3) a global research project (Fairwork) focused on 'cloudwork', which has included surveys with 613 workers in 84 countries (between 2020 and today) (Graham et al., 2020).

In the 'Microwork, crowdwork and AI data work' section, our analysis of the AI data pipeline draws more specifically on our fieldwork at an AI data BPO to provide insight into work practices through an example of the end-to-end AI data services that BPOs offer AI companies. We conducted this fieldwork at three delivery centres of the BPO in Nairobi, Kenya and Gulu, Uganda in April and May 2023. It consisted of workplace observations, presentations from management, and interviews with workers and managers ($N = 46$). This research enabled us to understand the most up-to-date labour processes and management techniques involved in this AI data work institution. We draw from this data to construct a more general model of the extended process AI companies must undertake, either themselves or with the assistance of external partners, to prepare and evaluate their datasets.

We draw on a specific type of data work institution that specialises in a sub-set of machine learning called computer vision, although the notion of an 'AI data pipeline' has similar formulations across the industry. We define an AI data pipeline as the set of data processing activities necessary to integrate datasets into the training and testing of machine learning models. We develop our own typology of the various stages of the AI data pipeline by drawing on industry sources and synthesising them into a conceptual framework that can help make sense of the role different AI data work institutions play in the overall process.

Microwork, crowdwork and AI data work

As the sophistication and scale of machine learning algorithms have increased, AI companies have a growing need for high-quality and low-cost sources of data (Bender et al., 2021; Crawford, 2021; Dauvergne, 2022). The production of this data requires a significant amount of human labour which includes the work of software developers and machine learning engineers who design and build AI systems, along with other data workers who are required to categorise, annotate and evaluate the data inputs and outputs of training programs (Miceli et al., 2022). 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. Our definition does not include software developers and machine learning engineers; nor does it include content moderators for social media platforms whose work does not feed into AI production.

Assignments of AI data workers consist of a variety of tasks, from categorising and assembling datasets, to annotating different types of data and interpreting and correcting

the results of machine learning algorithms (Miceli and Posada, 2022; Tubaro et al., 2020). Workers' actions and the organisational structure of their work can have broad ethical and political implications for how AI systems operate and their corresponding social effects on the world (Bender et al., 2021; Posada, 2021). This is through the often under-emphasised interpretive aspects of AI data work whereby social values and biases are embedded in data through both task design and the completion of tasks (Paullada et al., 2021).

We consider the term AI data work as a necessary addition to existing terminology because it provides a more precise and specific definition of this type of work that is not adequately captured by existing terms. In this section, we show why AI data work is to be preferred over related concepts. We follow Miceli and Posada (2022) in employing the term 'data work' as applying to the activity of curating, annotating and verifying datasets, but we seek to offer a more precise analysis of how this particular form of work applies to the production process of AI systems. Scholars in a diverse range of fields from healthcare to education have analysed the specific work activities of an emerging field of data occupations of those who produce and maintain datasets (Bossen et al., 2019; Lu et al., 2021). We contribute to these studies to show how AI systems have their own hidden labour that enables the more visible work of machine learning engineers training AI models. Unlike more general studies of data work in different industries, we specifically examine the AI production process. There is an overlap between our term and some of these other studies insofar as 'data work' undertaken in the healthcare industry might be for an AI system to be used by medical professionals in which case we could call this AL data work. Our study of AI data work builds on existing studies of data management outside of the AI industry that shed light on the 'backroom' work of organising datasets which feed into the 'front-end' work of data analytics (Parmiggiani et al., 2022; Pine and Bossen, 2020).

We also argue that AI data work is a preferable term to a host of other potential concepts already employed in the production of AI systems because it offers a more precise formulation of the work involved and reduces ambiguities in what is being referred to. For example, we employ the term AI data work rather than the broader category of microwork because the latter includes activities that are not related to the development of AI systems. Microworkers on digital platforms can undertake AI data work, but their tasks can also include consumer and academic surveys, translation tasks, providing feedback to companies on products, and a range of other tasks (Berg et al., 2018; Irani, 2015b). The term microwork was introduced by Jeff Bezos in 2006 when he was presenting Amazon Mechanical Turk (MTurk) at MIT: 'Think of it as microwork, so for a penny, you might pay someone to tell you if there is a human in a photo' (Jones, 2021a). The founder and then CEO of Sama, Leila Janah presented the concept as a way of providing digital work to underemployed populations in East Africa to help them overcome poverty and allow them to participate in the digital economy (Janah, 2017; TEDx, 2010). It has since been further developed to refer to a series of small tasks posted to on-demand labour platforms to be completed online by multiple workers (Irani, 2015b). In platform-based microwork, jobs are not offered to an identifiable subcontractor, but placed as an open call on a platform for any worker to fulfil. Microworkers are paid as little as a few cents per task and are classified as independent contractors without the benefits and protections of an employment contract (Berg et al., 2018).

Gray and Suri (2019) helped popularise the idea of microwork through their 2019 book, *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. They define ghost work as 'the human labor powering many mobile phone apps, websites, and artificial intelligence systems [which] can be hard to see' (2019: 7). This category of ghost work is broader than both AI data work and microwork as it includes many different varieties of online work – such as supporting companies with search results, performing translation tasks and verifying customers' accounts; it also includes much larger tasks, 'macrowork', that take workers a longer time to complete and require higher levels of skill and experience. The authors mainly studied distributed workers who performed tasks posted to platforms – both generalist crowdsourcing platforms such as MTurk, but also internal crowdsourcing platforms used exclusively by one tech company, such as Microsoft's Universal Human Resource System (UHRS). The essential aspect of this type of work for the authors was the oftentimes opaque employment practices that surround it and the fact that these systems tend to render workers' labour invisible in the development and operation of larger socio-technical systems. AI data work, in contrast, defines a much more limited and defined set of activities related to the preparation and verification of datasets in the production of AI systems. While there are significant overlaps with Gray and Suri's concept, tasks specifically referred to by the authors such as moderating social media content and answering a web-based customer chat query fall outside of the more defined scope of AI data work.

While microwork is a term that tends to be used for the types of short, often repetitive and low-skill online tasks that requesters post to platforms such as MTurk, 'crowdwork' and 'crowdsourcing' can include a wider variety of activities. Some authors use crowdwork as synonymous with microwork (Altenried, 2020), while others use it in a more expansive sense (De Stefano, 2016; Howcroft and Bergvall-Kåreborn, 2019). In their typology of crowdwork platforms, Howcroft and Bergvall-Kåreborn (2019) place 'online task crowdwork' or microwork alongside

'professional-based freelancing' on platforms such as Fiver, 'asset-based services' such as through Uber and Airbnb and 'playbour', unpaid creative work initiated by requesters. For these authors, 'crowdsourcing is highly heterogeneous and includes capital (crowdfunding), ideas generation (crowdsolving and competitions) and the polling of public opinion (crowdvoting), to name a few examples' (Howcroft and Bergvall-Kåreborn, 2019: 22). While most crowdsourced workers remain geographically isolated, this work can be socially embedded in online communities that assist workers in dealing with challenges related to their work (Schwartz, 2018). For the purposes of our typology of AI data work, crowdsourced work and crowdwork refers to a much broader field of activities and organisations than that which is relevant for an understanding of the labour integrated into AI systems.

Cloudwork refers to any work that is mediated via a digital labour platform and can be performed remotely by workers in a different location from the task provider. It is a term used to differentiate this form of platform-mediated work from 'tethered' or location-based work in which the workers must be in the same geographic location as the client or requester such as food delivery or ride-hail work (Lubke et al., 2023). This term is both too broad and too narrow to adequately capture AI data work. On the one hand, not all AI data work is performed remotely. Some of it is performed on the same premises either of the AI company itself or of an outsourced provider. On the other hand, it could refer to a wide range of digital tasks that have no relevance to AI systems such as 'translation, design, illustration, web development, and writing' (Lubke et al., 2023).

One of the key types of AI data work that tends to be left out of discussions is work performed not by geographically-dispersed independent contractors, but by employees of BPO centres, some of which are located in countries in the Global South such as the Philippines, India, Kenya and Venezuela. Miceli and Posada (2022) identified two different ways in which AI data workers can work: via the types of crowdsourcing platforms just discussed and specialised BPO companies. A BPO can be defined 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' (Miceli and Posada, 2022).

AI data workers at BPOs can be short-term or long-term employees of an organisation and can potentially have worked for long periods at the firm. BPOs can be more specialised than digital platforms and focus on specific types of data services and particular domains of application (such as autonomous vehicles or computer vision). They also tend to be more expensive than digital platforms because they can guarantee a higher quality of service with direct lines of communication between the client, management and workers. They also typically guarantee a level of information security that is unavailable through platforms. Many of the existing studies on AI data work consists of research on digital labour crowdsourcing platforms and independent contractors or 'microworkers' (Miceli and Posada, 2022; Tubaro and Casilli, 2019; Tubaro et al., 2020). Our article seeks to add to this important research on digital platforms by foregrounding the important role played by BPOs in the AI data industry.

Theorists have analysed the different functional roles that AI data work plays in relation to AI systems. Tubaro et al. (2020), for example, examined the role of digital platform labour and what they refer to as 'microwork'. They focus on the functions performed by microworkers recruited through digital platforms in three distinct forms of work: 'artificial intelligence preparation" "artificial intelligence verification" and "artificial intelligence impersonation' (Tubaro et al., 2020: 1). In each of these categories the authors show how the labour of microworkers is a crucial input to the production of AI systems; the work of machine learning engineers would not be possible without this labour. They argue that this form of labour is a structural component of the AI production process which is unlikely to be made autonomous soon as the technology reaches a more mature stage of development (Tubaro and Casilli, 2019; Tubaro et al., 2020). Our article seeks to build on these insights by undertaking a closer analysis not only of the functional role of data workers, but of the specific practices they perform and how these fit into the production of AI systems.

AI data workers' labour is embedded not only in larger production networks of AI systems, but also in 'planetary labour markets' in which tech companies are searching for the cheapest possible source of labour to fulfil their AI data work needs (Posada, 2021). As a result, much of this work takes place in different locations in the Global South, including Latin America, Asia, Africa and the Middle East (Jones, 2021b; Muldoon et al., 2023; Posada, 2021). In this article, we examine the different data work institutions in which this work is performed.

AI data work institutions

An AI data work institution is an organisation that arranges for AI data work to be undertaken either by employees of the organisation within a designated facility or by geographically-dispersed independent contractors. We follow W. Richard Scott (2013: 56) in defining institutions broadly as consisting of the 'regulative, normative, and cultural-cognitive elements that, together with associated activities and resources, provide stability and meaning to social life'. In particular, we emphasise the formalised and regulatory aspects of institutions that control individuals' and firms' behaviour in competitive markets. From this economic perspective, Douglas North (1991: 97) emphasises that institutions 'define the choice set and

Table 1. AI data work institutions.

			Type of worker	
			Crowdsourced/self-employed	*Employees*
Type of service	***External***	*Generalist company*	*Type A*: Generalist platform	*Type B*: Generalist BPO
		AI data company	*Type C*: AI data platform	*Type D*: AI data BPO
	Internal	*AI company*	*Type E:* Internal data platform	*Type F:* Internal data services

AI: artificial intelligence; BPO: business process outsourcing.

therefore determine transaction and production costs and hence the profitability and feasibility of engaging in economic activity'. We continue the application of this institutional perspective to the study of digital platforms undertaken by Niels van Doorn (2020) and Benjamin Bratton (2016) and focus more specifically on institutions that organise data work for the production of AI systems. Table 1 demonstrates how a wide variety of empirical cases could fit within certain ideal types that share key attributes (Weber, 1949).

The table synthesises existing literature on AI data work into a 3-by-2 table which generates six canonical institutions that arrange for this work to be performed. These types could be considered as an analytic tool that helps make sense of the messy empirical reality of AI data work and reduce it to a set number of core organisational types. We categorised these AI data work institutions by first conducting a systematic review of the literature to determine the broad range of institutions that facilitated work that would fall within our definition of 'AI data work' (including what other authors classified as microwork, cloudwork and other related concepts). We excluded any charitable or educational organisations and focussed exclusively on companies undertaking paid services as a business activity. Following this review, we distinguished relevant institutions that performed data work according to three questions, which resulted in six different categories. The first criterion for categorising data work institutions is related to the nature of employment relationships within the institution. Did the institution engage crowdsourced workers or did they employ geographically tethered workers that worked inside the institution's physical premises? Second, we inquired about the type of work undertaken within the institution. Was the institution exclusively performing what we defined as 'AI data work' or did it perform a variety of other functions such as office administration or other forms of microwork. Third, as an overarching point, we asked if the AI data work institution could be considered as an outsourced external partner to the organisation developing the AI system or whether it was based within the company. Following this procedure of categorising AI data work institutions, we produced the matrix shown below.

Type A: Generalist platform

Generalist microwork platforms such as Amazon Mechanical Turk and Microworkers operate online marketplaces for digital tasks which enable requesters to post a variety of jobs online to be performed by a geographically distributed workforce (Berg et al., 2018; Bergvall-Kåreborn and Howcroft, 2014; Casler et al., 2013). The close human management of workers is replaced by a digital system of verification that enables requesters to judge the quality of work and choose not to pay for poor-quality tasks (Irani, 2015a). Microworkers have been found to be a heterogeneous group of workers with different motivations for performing microwork and different levels of dependency on microwork platforms. Generalist platforms have been uncritically touted as a progressive potential for enabling new populations to secure work for workers who would otherwise find it difficult to access traditional labour markets (Gupta, 2017). However, researchers and regulators have expressed a growing concern for the precarious working conditions, the lack of minimum wage protections and the frequent non-payment of tasks on these platforms (Aloisi and De Stefano, 2022; Chen et al., 2019). Regulating these platforms has proved difficult because of the geographically dispersed nature of the workforce, the lack of fit within existing legal frameworks, and because governments in low-income countries may wish to encourage foreign investment into the BPO sector (Berg et al., 2018).

Requesters with AI data work requirements can post jobs to the platform to have them performed by workers as individual tasks. The advantage of this approach compared to having their own employees undertake this work is that companies can access a large workforce at a relatively cheap cost who can perform the large quantity of work necessary to categorise and annotate data if it is broken down into short, distinct tasks. Large platforms with thousands of workers have been one important way in which AI companies have been able to undertake data work, but there are also disadvantages of this option. First, our fieldwork suggests that workers rarely have special training on the type of data work the company requires undertaken. Workers on generalist platforms typically perform a wide variety of tasks in addition to AI data work such as filling in surveys and performing online searches for companies.

These workers do not receive any specialist training on particular requirements for AI systems such as how to annotate a busy street scene for autonomous vehicle software. This results in a lower level of quality and specialisation compared to a more consistently skilled workforce, as will be described below. Second, the assessment of data quality may require either specialised knowledge to understand what constitutes ground truth data, or technical skill required to compute automated quality metrics such as the agreement between workers on tasks. Consequently, the AI company may end up owning the expensive task of quality control with little ability to input in the quality of data produced by the AI data workers. Third, work typically cannot be sent back to the platform if it is poor quality which may adversely affect the accuracy and quality of algorithmic outputs.

To take one example, in the case of companies that develop autonomous vehicle systems, their algorithms require an extremely high degree of accuracy because errors in determining the movement of a vehicle could prove fatal (Tubaro and Casilli, 2019). The security and privacy of sensitive data, particularly personally identifying information, can also be an issue on large platforms. This is due to the relative lack of oversight on data accessed via thousands of distributed workers, compared to being accessed by verified employees of a secure BPO facility who are under non-disclosure agreements (NDAs). Platforms typically do not offer end-to-end services when it comes to processing and managing data. AI companies may have to prepare their own datasets and design tasks based on the platform's parameters, including the development of training and instructions for workers. These factors increase the competitiveness of data work BPOs who may have more secure privacy and security protocols to accommodate for the end-to-end management of complex and sensitive data projects.

Type B: Generalist BPO

AI companies can choose to outsource AI data work to BPO companies. In this process, particular aspects of the AI data work are delegated to an external provider who manages the tasks based on defined performance metrics set by the AI company. Generalist BPOs are not specialists in AI data work; they offer a wide range of services to clients. IT-based BPO work grew rapidly in the 1990s and 2000s due to the spread of the Internet and ICT services globally, which have reduced communication costs and enabled new international partnerships (Lacity et al., 2011). The BPO industry offers a variety of services such as finance, logistics and HR; it also offers domain-specific forms of specialisation such as in healthcare, retail and banking (Mehta et al., 2006). BPO work can be outsourced domestically to a vendor in the same country or internationally to an overseas vendor. Some of the largest BPOs offer comprehensive services in which the vendor takes responsibility for multiple parts of a company's business. In certain cases, businesses outsource all of their back-office processes to vendors in order to take advantage of the reduced labour costs and the competitive advantage that companies have based in locations such as India or the Philippines (Mehta et al., 2006).

A BPO offers clients several distinct advantages over crowdsourced platforms. First, the management structure and forms of labour control available at BPO delivery centres allow for closer supervision of workers which can result in a higher quality of outputs. Detailed instructions on how to complete the tasks can be sent to the BPO vendor which can be passed down the chain of command to the data workers whose processes can be monitored digitally and by human quality assurance agents. If those quality assurance agents notice 'edge cases' that go beyond the instructions, then the existence of these can be fed back up the chain to the client and the instructions adjusted accordingly. Second, the structure of BPOs enables them to offer specialised end-to-end services in which an AI company's data is managed on the BPO's platform. Rather than having small discrete tasks distributed across a large workforce, BPOs can manage several stages of a client's data needs, which increases the quality of the outputs and allows for better management of the data at each stage of the process. Third, dealing with a single supplier provides opportunities for an AI company to provide continuous feedback to the vendor to improve the process. Batches of work can be sent back for revision and instructions can be provided on the precise aspects of the data work that need to be changed. The vendor retains an institutional knowledge of the workflow and can update their methods to meet the client's needs. Finally, an AI company can engage the services of multiple BPOs to perform different aspects of their data work and establish incentive structures in which the performance of each BPO is monitored and ranked against their competitors to induce higher levels of performance through rewards for the top-performing vendors.

Type C: AI data platform

Workers on AI data platforms are neither employees of the AI company nor the platform. Instead, workers are independent contractors who are geographically distributed, sometimes based in a particular region but often across the globe. As a data requester, an AI company can post tasks to be completed by individual workers, which are then returned individually without any coordination between the workers. The growth of the AI data work sector has led to the emergence of new competitors to traditional platforms like MTurk; these competitors offer more specialised services to support AI systems. Schmidt (2019) highlights that many AI companies with complex needs and

requirements for a high degree of quality are increasingly moving towards more specialised services. These clients may have more involved tasks that need a higher degree of patience and skill to complete, requiring workers to review long videos to tag specific events or to undertake semantic segmentation on images involving a long process of identifying a vast number of different objects and annotating them correctly.

AI data platforms may require annotators to complete training courses before being allowed to access specific kinds of more skilled tasks such as Light Detection and Ranging (hereafter LiDAR) a method for determining ranges through laser imaging which is a form of 3D laser scanning that is used for 3D moving objects. This drives clients to seek more expensive and specialised services that emphasise a higher degree of workforce management and quality control. Schmidt (2019) argues that younger AI data companies with a growing client base such as Mighty AI, Hive AI and Playment are often preferable for clients that appreciate the speed and low cost of traditional crowdsourcing platforms, but require higher precision for their AI models because small errors in the training data make these models less effective.

Tubaro and Casilli (2019) noted that some platforms such as Appen and IsAHit explicitly market themselves to specific industries such as automotive companies developing autonomous vehicles. We also find companies seeking to specialise in one variety of AI services such as Sama increasingly moving towards computer vision AI data work. Platforms also specialise in particular roles within the establishment of two-sided markets for data work, for example, only handling the recruitment of workers, the allocation of tasks or managing the interface where work is performed (Tubaro and Casilli, 2019). Some clients with complex requirements value these services because they offer a greater degree of specialisation and provide a layer of opacity that make it more difficult for competitors to determine the precise nature of commercial relationships.

Type D: AI data BPOs

Just as the need for greater levels of specialisation in AI systems creates a market for AI data platforms, so too have specialised BPOs emerged to cater to client's complex AI data work requirements. Examples include Cloudfactory and Sama, based in Nepal and East Africa respectively. The advantage of these BPOs is that they can train large workforces in particular types of AI data work, such as image, video and 3D moving object annotation. Additionally, they can build expertise in specific industries, subfields of machine learning and types of AI data tasks; providing them with a competitive advantage compared to generalist platforms or BPOs. Our fieldwork at Sama revealed that the company has detailed records of which workers have specialisation in particular tasks, which allowed them to quickly ramp up large projects with highly-trained workers. When workers complete a specific project for a client this would be stored on their record. Sama classified different skills based on how long it would take workers to become proficient in a task such as annotating street scenes, identifying objects a smart vacuum might encounter in a family home, and LiDAR, the latter being among the most complex of tasks workers could be trained in. Many workers at Sama also had long periods of employment with the organisation, with a large percentage of the workers having worked there for at least three to five years.

Specialisation in AI systems also provides an opportunity for BPOs to develop their own platform and services for managing AI data, which reduces friction between the different stages of the AI data pipeline. Sama workers were trained in how to use the SamaHub platform, proprietary software that enabled workers to engage in customised data annotation projects. Clients had the option of either using SamaHub or allowing Sama workers to access their own AI platforms to complete tasks (Sama, 2023b). If clients opt for an AI data BPO's proprietary platform, they also benefit from integrated tools that augment the efficiency of human AI data work through automated annotation, curation and evaluation. Sama claims that the ML-assisted AI data pipeline results in a 2-4× increase in efficiency. Similarly, Cloudfactory (2023) advertises that its platform can 'accelerate the AI Lifecycle with Human-in-the-loop Solutions' and 'deliver accurate labels 5x faster with AI-assisted labeling and fully integrated humans in the loop'.

Type E: Internal data platform

Several of the large tech companies use their own internal data work platforms, which have usually been modelled off the operations of MTurk (Gray and Suri, 2019). Crowdsourced workers on these platforms can be requested to perform a variety of different tasks for employees of the company to improve internal products and proprietary algorithms. Instead of posting a task on a generalist crowdsourcing platform, company employees can utilise their own platform to ensure that workers have signed non-disclosure agreements, which maintains higher levels of privacy and security for the company's products. Microsoft's internal company platform is called Universal Human Resource System (UHRS), and its website states that it is 'a crowdsourcing platform that supports data labeling for various AI application scenarios. Our vendor partners connect us with people – who we refer to as "judges" – to provide data labeling for us at scale' (UHRS, 2023). Microsoft's employees and authorised partners are the only people allowed to submit a request on UHRS.

The platform offers a range of services in relation to content moderation, image annotation and video annotation

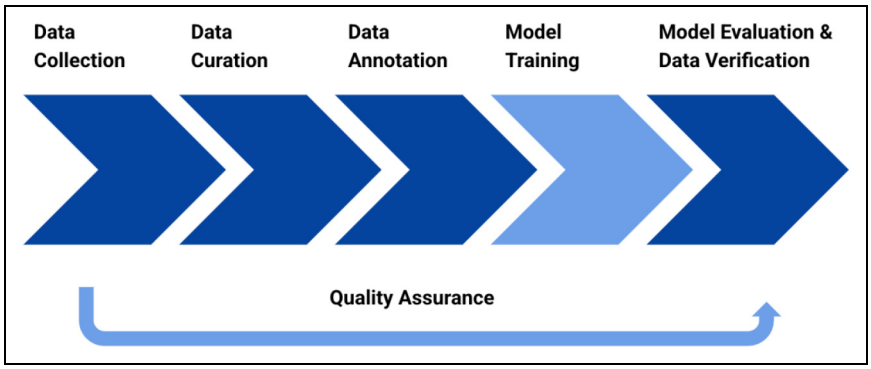


Figure 1. The artificial intelligence (AI) data pipeline.

including the ability to 'classify thousands of images quickly and easily. Train your products and tools with improved image detection, boundary recognition, and more with high-quality annotated image data' (UHRS, 2023). Interviews with workers by Gray and Suri (2019) indicated that workers also review voice recordings, rate the sound quality of clips and moderate text for adult content. Workers who perform tasks on these platforms are not employees of the tech companies. They are recruited through a third-party 'vendor management system' (VMS) who organise workers and make sure they sign NDAs. For example, Google uses a VMS to find workers for its microwork platform, while Twitter and Facebook use internal platforms that function in a similar way to MTurk to source workers to monitor and review content on their respective platforms (Gray and Suri, 2019).

Type F: Internal data service

AI companies may also employ in-house workforces to perform AI data work which can include full-time employees and contractors (Partnership on AI, 2021). This approach allows AI companies to build closer relationships with in-house workforces who may be located on the same premises as the broader project team. In-house AI data workers are also able to develop more nuanced institutional and technical knowledge which can improve the quality of the data annotation process. The downside to this system is that AI data work is extremely labour-intensive and paying full-time employees based in high-wage jurisdictions such as the US and Europe is more expensive than outsourcing this work to a data work institution in a jurisdiction with comparatively lower wages. If a company uses its own machine learning engineers to prepare datasets this also takes highly trained personnel away from projects that require more training and experience. Consequently, AI companies often combine internal data services with other AI data work institutions such as external platforms or BPOs, depending on project constraints such as budget or the scale and quality of data required (Cavello, 2020). For example, an internal team may run a pilot with a smaller subset of data to enrich understanding of the project's data needs, or refine task design for an external data work project (Partnership on AI, 2021).

The AI data pipeline

AI data workers are required for a variety of different tasks in the AI production process, from very early stages of data collection and organisation, up to the final stages of model evaluation and data verification. Rather than conceptualise these tasks as discrete moments, it is more useful to see them as interconnected parts of a single process. How data is managed at one stage of the process can have significant impacts upon later stages; poorly curated data may magnify biases or inaccuracies in later stages. We refer to this as the AI data pipeline (see Figure 1) and track the continuities between different stages of this process. It is worth prefacing that the pipeline is not designed to be rigid but represents an idealised workflow to provide specificity about how AI data work is integrated into AI production processes. In practice, the pipeline is not always linear and many stages of the process are fragmented, iterative or modular, depending on the requirements of different data projects. For example, some projects may focus on annotating images in a large-scale computer vision dataset, while the annotation of LiDAR data for an autonomous vehicle application may require significant back and forth between the AI company and the AI data company to ensure high accuracy. Our pipeline primarily draws from case studies of AI data work for computer vision and associated applications such as autonomous vehicles, logistics and retail; the pipeline would likely differ for algorithms across different modalities such as language or audio.

AI data work institutions may manage all of these stages for a client including partial automation of stages through a digital platform and machine learning tools (Partnership on AI, 2021). In our fieldwork, this entire process could be managed on the company's bespoke platform, SamaHub.

Sama (2023a) advertises to clients that 'we provide data scientists, ML engineers, and data operations teams an integrated platform for AI data preparation, labeling, and collection'. As we have stated, SamaHub is integrated with machine learning tools that automate some aspects of data curation, annotation and validation. This type of service would be difficult for a digital platform with a distributed workforce to offer due to the decreased labour control and lack of an integrated workflow between different stages. This pipeline consists of several important stages, which are conceptualised in Figure 1. This conceptualisation allows for important distinctions to be made between different aspects of the preparation of AI datasets such as 'data collection', 'data curation' and 'data annotation' which are sometimes either blended together or confused in existing research (Tubaro et al., 2020).

The AI data pipeline also helps conceptualise many of the advantages of certain types of AI data work institutions as described above. As we have stated, one of the limitations of generalist platforms is that they cannot offer the kind of end-to-end services that more specialised AI data BPOs can offer. AI companies increasingly have complex data needs and require more specialised services that can assist them with multiple stages along the AI data pipeline. Platform-based data annotation work distributed to multiple workers across the globe such as in generalist platforms and AI data platforms has limited opportunities for quality assurance and iterative cycles of annotation and model improvement. Platform-based AI data work tends to be a specific input into a pipeline organised by the client; platforms generally do not take on a larger responsibility for multiple sequential steps in the pipeline as a whole. Understanding the complexities of the pipeline therefore provides a new vantage point from which to assess the adequacy of different AI data work institutions for the needs of contemporary AI companies. The pipeline itself was developed through fieldwork at an AI data BPO and demonstrates some of the advantages of using this type of institution for complex data handling needs.

Data collection

Data collection involves the acquisition of existing data from third-party sources or the creation of new data via human data collection. AI companies often purchase datasets from third-party providers or compile their own from multiple sources, including open-source repositories, proprietary licenses and via web scraping (Gebru et al., 2021). On occasion, these companies require additional data, which is a role that can be performed by AI data workers to help compile and create datasets. This is not a core feature of data work BPOs since many AI companies come with their own sources of data which they require to be organised and annotated. However, at Sama we documented teams of workers performing web-based searches for clients, determining if particular products were being sold in specific markets and compiling a database of evidence. Sama does not advertise data collection on their website as one of the main services they offers clients, but it was clear from worker interviews that a small amount of the work undertaken within the company would fall within the sphere of data collection. Tubaro et al. (2020: 5) also provide examples of platform-based microworkers generating data for clients such as audio utterance collection in which 'platforms can leverage their contributor base to gather this data with a variety of vocal timbres, regional accents, uses of slang and contexts'. Data collection services can include workers themselves generating data through creating text, audio and video files in addition to workers performing research and creating datasets by adding their work to existing datasets.

Data curation

Data curation consists of tasks for processing data through editing, filtering and analytics. These improve the efficiency and reduce the cost of downstream AI data tasks through the identification of the data most likely to improve a model's accuracy and performance. For example, data curation for an autonomous vehicle algorithm might involve preprocessing image and video data to identify objects that have good light, clarity and positions to be recognised, which can make production data processing more efficient. This process occurs before data annotation commences, but it can also be iteratively undertaken as part of a continuous feedback loop to help optimise the efficiency of data annotation processes. This stage of the AI data pipeline is often missed in the current literature on data workers because it is not typically performed by platform-based microworkers. The advantage of BPOs over a digital labour platform is they can offer more complex and integrated services which are more efficient and ensure a higher quality of outputs.

Data curation can assist AI companies with a number of distinct problems they face with leveraging their existing datasets. Often companies have too much data to be cheaply and efficiently annotated and must decide how to filter and organise it. When a dataset contains large amounts of uninteresting data and small events that provide important source material for machine learning models, data curation can help isolate these moments so that only the richest information is offered to annotators. Companies might also not know how to choose data for annotation and might engage in the labour-intensive and costly process of manually picking data to be sent to data annotators. Data curation helps AI companies select which data is most likely to improve the performance of a model from the larger unlabelled dataset. Even if AI companies know which data they would like to annotate, it can often be difficult for them to organise their datasets

and extract the desired files. Data curation tools can offer automated or semi-automated processes to filter data based on a company's search criteria.

Data annotation

Data annotation is the largest and most time-consuming part of the AI data pipeline, and comprises the core service that data work institutions offer their clients. This stage of the data pipeline can take a variety of different forms depending on the needs of the client and the type of data requested. Sama (2023a) specialises in computer vision and so many of the data annotation tasks at the company consisted of the tagging image, video and LiDAR data. Tasks we observed included bounding boxes (drawing rectangles around desired objects), polygons (multi-sided shapes that tightly follow the outline of an object), keypoints (detects pose variations for motion tracking, facial landmark detection, and hand gesture recognition), semantic segmentation (tags precise edges of objects to differentiate between different areas in an image or video) and lines and arrows (directional indicators). Many of the projects undertaken at Sama are performed on the SamaHub platform, which allows workers to rapidly create high-quality annotations on client data. Workers are trained in how to use the tool, which also comes with simple instructions and pop-up boxes that explain key processes if workers require more assistance. Developers within the company routinely update the tool and show workers how to use new features.

Clients had different quality thresholds for different projects, but all projects at the company had to aim for a minimum 95% accuracy score which would be assessed by quality analysts within the company and by audits from clients. For certain projects, particularly those concerning autonomous vehicles, clients would require an even higher level of accuracy, sometimes up to 98%–99%, which would create extra pressure on annotators and the quality assurance team. Some of these clients also required much more detailed forms of annotation with pixel-accurate identification of objects rather than the rougher bounding boxes that would indicate the general position of an object without the additional accuracy.

Quality assurance

A high quality of outputs must be maintained at every stage of the data pipeline to ensure the accurate functioning of machine learning algorithms. Quality assurance consists of both manual and automated tasks. Manual tasks include random or isolated sampling of annotated data to check quality across certain classes of data or reviewing disagreements between multiple data workers on a task. Automated quality assurance might include algorithmic tools to drive analysis of outputs or comparison of data to high-quality 'ground truth' data completed by human experts (Krig, 2014). In the case of an autonomous vehicle case study, this might involve checking for incorrect labels, missed objects or points, or misinterpreting an object's size or direction of travel (Walker and Steves, 2023). AI data companies usually have dedicated quality assurance teams that check the quality of the annotated data and make sure it meets the quality thresholds agreed upon by clients. Sama had a dedicated team of quality analysts who were monitored by quality assurance supervisors double-checking work undertaken by data annotators. These employees were often former annotators who were identified as particularly accurate in their work and able to provide support to other employees in improving their quality. They were considered more senior in the company than annotators and received a small increase in their pay to reflect their increased responsibilities. In one example we observed, a quality assurance supervisor was checking the annotation of a street scene and determining whether every object in the image had been correctly annotated. Quality assurance would occur as an ongoing process during the annotation of datasets and also at regular weekly performance monitoring assessments. Clients could also regularly audit work and would send work back to be redone if it did not meet their expectations for quality. Sama's quality assurance team would proactively raise edge cases with the client and seek further advice in cases of ambiguity where the correct annotation could not be adequately determined.

Model training

Once the data annotation work is completed and all quality checks have been made, the processed data is then returned to the client to be used in machine learning algorithms. AI data work institutions tend not to be involved in this stage of the process; AI companies typically hire specialised practitioners such as machine learning or software engineers, and data scientists to design and deploy models, including project managers or specific operational staff to manage relationships with AI data work institutions (Partnership on AI, 2021). Once the pre-processed data has been returned to the team, the training data is ingested into the machine learning model which identifies patterns in the data that allows a production software to compute important predictions. In the case of an autonomous vehicle, this might involve the identification and categorisation of objects on the road such as other vehicles, cars and pedestrians (Sama, 2023b). This process allows the algorithm to learn from examples in the training data which are enriched by the annotations provided by AI data workers during the pre-processing stages of the AI data pipeline. During production, the model might encounter data that is unlike the training data; this is called data drift and results in low accuracy and model performance. Understanding the models' accuracy and performance is referred to as Model

Evaluation, which can be conducted both within AI companies and as an outsourced service provided by AI data companies.

Model evaluation and data verification

Model evaluation and data verification are processes to analyse the accuracy and performance of AI systems, often to inform an iterative process of preparing new data. During the model evaluation, model accuracy and performance are measured by either comparing aggregate predictions to a test set consisting of 'gold tasks' or comparing them to other models to calculate an accuracy score. In the case of an AV system, this might relate to evaluating for performance on object detection and tracking tasks. This process may result in AI data workers '[producing] training data from the amended outputs of an already-trained algorithm' (Tubaro et al., 2020: 8). Tubaro et al. (2020: 8) referred to model evaluation as 'AI verification' in which data workers check the accuracy and quality of algorithmic solutions.

Data verification on the other hand refers to reviewing and correcting model predictions, or labels in the training data. This might involve identifying errors related to objects that are underrepresented in the training data in diversity and quantity. This process is difficult to measure in aggregate as human judgement is often required for complex or subtle nuances in data. For example, annotating the perceived fatigue of a driver for an AV 'In-cabin Behaviour Monitoring' solution. Tubaro et al. (2020: 8) noted that microwork platforms 'advertise services such as relevance scoring and transcription checking to their AI-producing clients' but do not specifically sell these services as data verification.

Conclusion

This article presents a typology of AI data work that adds nuance to existing conceptual divisions about data work institutions and different stages of the AI data pipeline. We draw from a wide range of fieldwork, including research conducted at a BPO in Kenya and Uganda that focuses on the end-to-end delivery of AI data work for computer vision algorithms and associated retail, automotive and logistics applications (Muldoon et al., 2023). The AI data industry is likely to grow alongside the increasing importance of machine learning algorithms for a range of industries including transportation, retail and healthcare. Indeed, the nature of machine learning requires the ongoing acquisition of new data to ensure that algorithms can function in real-world production settings (Tubaro et al., 2020). Although many aspects of AI data work can be augmented by ML-assisted tools, the dynamic and complex realities of the real world mean that the human interpretation intrinsic to AI data work is likely to remain an important aspect of the AI data pipeline (Paullada et al., 2021).

The typology adds nuance to the relationship between workers and their experience, and the business strategies of AI data work institutions and their clients within broader AI production networks. AI data work is conducted and managed across different employment structures and spatial distributions, which are also shaped by the specific histories and cultures of the geographic locations in which the work is performed. However, across all typologies, concerns about fairness and labour exploitation have been identified: from the low ratings of crowdsourced platforms to the recent media attention on Sama regarding labour exploitation and union busting (GPAI, 2022; Perrigo, 2023).

In this article, we have focused on the organisational dynamics of the different employment structures and institutional forms of AI data work. Often these jobs are outsourced to the Global South and tend to consist of low-paid, monotonous and repetitive tasks that can be performed by workers without significant training or experience (Mohamad et al., 2020; Muldoon and Wu, 2023). When scholars refer to this type of work intermittently as either 'microwork', 'ghost work', 'crowdwork' or 'cloudwork' they can neglect the organisational dimensions of the type of labour performed and the specific working conditions of these workers. We have proposed the term 'AI data work' as a more precise term and have delineated the different types of institutions within which this work is performed. Through the use of this term and the corresponding institutional matrix, finer distinctions can be made when referring to how AI systems are produced and the different forms of labour that make this possible. The term allows for the distinctive elements of the AI production process to be differentiated from other forms of data work (say in education, healthcare or finance) in addition to showing how not all work on digital platforms (i.e. microwork or cloudwork) contributes to AI systems. Such distinctions should be important to policymakers, labour organisers and researchers because they determine the types of regulations that would be important to protect the relevant workers and the different forms of labour organisation that could be used to struggle for improved conditions.

To mention just one example, in the case of most platform-based microwork, workers are geographically distributed making it difficult to organise collective in-person protests, whereas AI data workers at BPOs can engage in more traditional strikes, pickets and collective action at their places of work. All of the tasks associated with the AI data pipeline can be performed by workers located within AI data BPOs, thus centralising the workforce and providing a potential lever for acts of collective resistance. This provides one crucial avenue for AI data workers seeking more equitable and just forms of work in the AI production process.

Acknowledgements

The authors would like to thank two reviewers for their positive feedback which has improved this manuscript. The authors are grateful to Jackie Kay and Maribeth Rauh for their valuable comments and suggestions on a draft of this article.

Declaration of conflicting interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of thisarticle: The German Federal Ministry for Economic Cooperation and Development (BMZ), commissioned by the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ), the Economic and Social Research Council through the Global Challenges Research Fund ES/S00081X/1), and the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013) [grant number 335716] for funding much of the research that informed this piece.

ORCID iD

James Muldoon https://orcid.org/0000-0003-3307-1318

Note

1. A full account of the methodology of this fieldwork can be consulted at James Muldoon, Callum Cant, Mark Graham and Funda Ustek Spilda, 'The Poverty of Ethical AI: impact sourcing and AI supply chains'. *AI & Society* (2023) doi.org/10.1007/s00146-023-01824-9.

References

Aloisi A and De Stefano V (2022) *Your Boss Is an Algorithm Artificial Intelligence, Platform Work and Labour*. London: Bloomsbury. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/bjir.12727 (accessed 23 February 2023).

Altenried M (2020) The platform as factory: Crowdwork and the hidden labour behind artificial intelligence. *Capital & Class* 44(2): 145–158.

Anwar MA and Graham M (2022) *The Digital Continent: Placing Africa in Planetary Networks of Work*. Oxford: Oxford University Press.

Bender EM, Gebru T, McMillan-Major A, et al. (2021) On the dangers of stochastic parrots: Can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event Canada, 3 March 2021*. ACM, 610–623.

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 3 September 2021).

Bergvall-Kåreborn B and Howcroft D (2014) Amazon Mechanical Turk and the commodification of labour. *New Technology, Work and Employment* 29(3): 213–223.

Bossen C, Pine K, Cabitza F, et al. (2019) Data work in health care: An introduction. *Health Informatics Journal* 25(3): 465–474.

Bratton BH (2016) *The Stack: On Software and Sovereignty*. Cambridge, MA: MIT Press.

Casler K, Bickel L and Hackett E (2013) Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior* 29(6): 2156–2160.

Cavello B (2020) Developing Guidance for Responsible Data Enrichment Sourcing. Available at: https://partnershiponai.org/developing-guidance-for-responsible-data-enrichment-sourcing/ (accessed 11 June 2023).

Chen W-C, Suri S and Gray ML (2019) More than money: Correlation among worker demographics, motivations, and participation in online labor market. *Proceedings of the International AAAI Conference on Web and Social Media* 13: 134–145.

Cloudfactory (2023) Data Labeling Solutions for Leading AI Innovators. Available at: https://www.cloudfactory.com/ (accessed 5 May 2023).

Cognilytica Research (2019) *Data Engineering, Preparation, and Labeling for AI* 2019. Available at: https://www.cloudfactory.com/reports/data-engineering-preparation-labeling-for-ai (accessed 4 May 2023).

Crawford K (2021) *Atlas of AI: Power, Politics and the Planetary Costs of Artificial Intelligence*. Yale and London: Yale University Press.

Dauvergne P (2022) Is artificial intelligence greening global supply chains? Exposing the political economy of environmental costs. *Review of International Political Economy* 29(3): 696–718.

De Stefano V (2016) The rise of the "just-in time workforce": On demand work, crowdwork, and labor protection in the "gig economy". *Comparative Labor law and Policy Journal* 37(3): 461–471.

Gebru T, Morgenstern J, Vecchione B, et al. (2021) Datasheets for Datasets. arXiv:1803.09010. *arXiv*.

GPAI (2022) *AI for Fair Work: AI for Fair Work Report*. GPAI.

Graham M (2015) Contradictory connectivity: Spatial imaginaries and techno-mediated positionalities in Kenya's outsourcing sector. *Environment and Planning A* 47: 867–883.

Graham M, Hjorth I and Lehdonvirta V (2017) Digital labour and development: Impacts of global digital labour platforms and the gig economy on worker livelihoods. *Transfer: European Review of Labour and Research* 23(2): 135–162.

Graham M, Woodcock J, Heeks R, et al. (2020) The fairwork foundation: Strategies for improving platform work in a global context. *Geoforum; Journal of Physical, Human, and Regional Geosciences* 112: 100–103.

Grand View Research (2022) *Data Collection And Labeling Market Size & Share Report* 2030. Available at: https://www.grandviewresearch.com/industry-analysis/data-collection-labeling-market (accessed 4 May 2023).

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

Gupta N (2017) *An ethnographic study of crowdwork via Amazon Mechanical Turk in India*. Ph.D. University of Nottingham. Available at: http://eprints.nottingham.ac.uk/41062/ (accessed 22 March 2021).

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

Irani L (2015a) Justice for "Data Janitors". In: *Public Books*. Available at: https://www.publicbooks.org/justice-for-data-janitors/ (accessed 3 February 2023).

Irani L (2015b) The cultural work of microwork. *New Media & Society* 17(5): 720–739.

Janah L (2017) *Give Work: Reversing Poverty One Job at a Time*. New York: Penguin Random House. Available at: https://www.penguinrandomhouse.com/books/546305/give-work-by-leila-janah/ (accessed 9 May 2023).

Jones P (2021a) Big tech's push for automation hides the grim reality of 'microwork'. *The Guardian*, 27 October. Available at: https://www.theguardian.com/commentisfree/2021/oct/27/big-techs-push-for-automation-hides-the-grim-reality-of-micro work (accessed 1 May 2023).

Jones P (2021b) *Work without the Worker*. London: Verso.

Krig S (2014) Ground truth data, content, metrics, and analysis. In: Krig S (eds) *Computer Vision Metrics: Survey, Taxonomy, and Analysis*. Berkeley, CA: Apress, 283–311.

Lacity MC, Solomon S, Yan A, et al. (2011) Business process outsourcing studies: A critical review and research directions. *Journal of Information Technology* 26(4): 221–258.

Lu AJ, Dillahunt TR, Marcu G, et al. (2021) Data work in education: Enacting and negotiating care and control in teachers' use of data-driven classroom surveillance technology. *Proceedings of the ACM on Human-Computer Interaction 5, CSCWS*: 1–26. DOI/10.1145/3479596.

Lubke G, Bhargava S, Valente J, et al. (2023) *Fairwork Cloudwork Ratings 2023: Work in the Planetary Labour Market*. Fairwork: Oxford Internet Institute. https://fair.work/wp-content/uploads/sites/17/2023/07/Fairwork-Cloudwork-Ratings-2023-Red.pdf.

Mehta A, Armenakis A, Mehta N, et al. (2006) Challenges and opportunities of business process outsourcing in India. *Journal of Labor Research* 27(3): 323–338.

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

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

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*. DOI: 10.1007/s00146-023-01824-9.

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

North DC (1991) Institutions. *Journal of Economic Perspectives* 5(1): 97–112.

Parmiggiani E, Østerlie T and Almklov PG (2022) In the backrooms of data science. *Journal of the Association for Information Systems* 23(1): 139–164.

Partnership on AI (2021) *Responsible Sourcing of Data Enrichment Services*. 16 June. Available at: https://partnershiponai.org/paper/responsible-sourcing-considerations/ (accessed 17 May 2023).

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 3 February 2023).

Pine K and Bossen C (2020) Good organizational reasons for better medical records: The data work of clinical documentation integrity specialists. *Big Data & Society* 7(2). doi.org/10.1177/2053951720965616.

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

Sama (2023a) Sama. Available at: https://www.sama.com/ (accessed 28 April 2023).

Sama (2023b) What is a Training Data in Machine Learning? In: SAMA. Available at: https://www.sama.com/training-data-inmachine- learning/ (accessed 11 June 2023)..

Schmidt F (2019) *Crowdproduktion von trainingsdaten: Zur Rolle von Online-Arbeit beim Trainieren autonomer Fahrzeuge*. Berlin: Hans-Böckler Stiftung. Available at: https://www.researchgate.net/publication/332964049.

Schwarz D (2018) Embedded in the crowd: Creative freelancers, crowdsourced work, and occupational community. *Work and Occupations* 45(3): 247–282.

Scott WR (2013) *Institutions and Organizations: Ideas, Interests, and Identities*, 4th ed. Los Angeles: Sage Publications.

TEDx (2010) TEDX Talk: Leila Chirayath Janah - 12/12/09. Available at: https://www.youtube.com/watch?v=1Ce9EfF2lHE (accessed 1 May 2023).

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): 205395172091977.

UHRS (2023) Universal Human Relevance System. Available at: https://prod.uhrs.playmsn.com/uhrs/ (accessed 1 May 2023).

van Doorn N (2020) A new institution on the block: On platform urbanism and Airbnb citizenship. *New Media & Society* 22(10): 1808–1826.

Walker R and Steves P (2023) Quality Ground Truth Labels for All Autonomous Driving Applications. In: *SAMA*. Available at: https://www.sama.com/ebook/quality-ground-truth-labels-for-autonomous-driving-applications/ (accessed 11 June 2023).

Weber M (1949) *The Methodology of the Social Sciences*. Free Press.