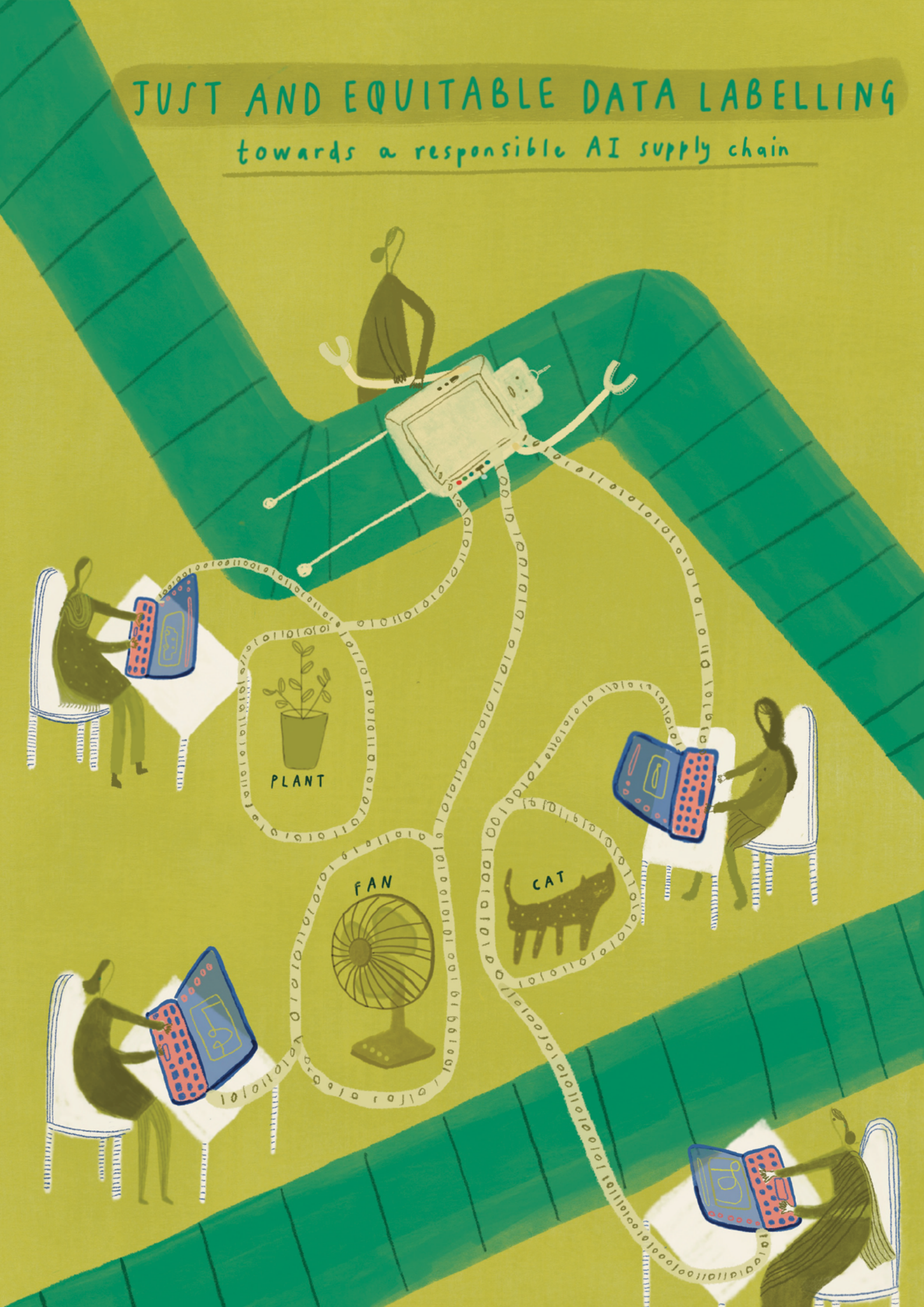
JUST AND EQUITABLE DATA LABELLING

towards a responsible AI supply chain



aapti institute

Dr. Sarayu Natarajan
Kushang Mishra
Suha Mohamed
Dr. Alex Taylor

Contents

1

Introduction · 01

2

Methodology · 04

3

Literature Review · 06

4

Findings from Primary Research · 09

4.1 Crowdwork to In-house work: Emergent Business Models in Data labelling
4.2 Safety, inclusivity, and well-being: practices in the industry Data Labelling
4.3 Evolution of AI data labelling work and emerging trajectories

5

Balancing employment generation with worker protection: policy pathways · 19

5.1 Considerations in policy-making
5.2 Potential Policy Pathways

6

Conclusion · 23

7

References · 24

8

Appendices · 27

8.1 Startups Interviewed
8.2 Interview Questions

1

Introduction

Jobs, worker well-being and Responsible AI

The Covid-19 pandemic has accelerated adoption of cloud-based services, Artificial Intelligence (AI) and Internet of Things (IoT) to further digitise and automate sectors like healthcare, education and entertainment (Chakravarty, 2020; Das, 2020). Emerging technologies like AI, however, are dependent on the availability of labelled and classified data sets, which must be tagged and annotated by hand. This work historically has been dependent on the gig economy and carried out by independent contractors on crowd-work platforms like Amazon Mechanical Turk, Appen or Clickworker (Gray and Suri, 2019). Increasingly, however, there has been a rise in private entities dedicated to providing labelling and annotation services. They serve global clients like Microsoft, TripAdvisor, eBay and Autodesk as well as provide employment to thousands of people (Joshi, 2019; Thomas, 2020). According to Cognilytica Research (2020), the market for third-party data labelling solutions is projected to grow to $4.1 billion by 2024.

This surge in automation and the necessity for a 'human-in-the-loop' for creating robust, training data sets is indicative of the fact that data labelling is likely to be a viable employment opportunity in India, particularly given that it can be carried out remotely. Its potential was also recognised by India's first-ever National Strategy for AI, by NITI Aayog (2018). The report acknowledged the importance of annotated data sets in the value chain of AI, the maturity of India's data annotation market, and its importance in generating significant employment. It also recommended the creation of an AI marketplace to support the startup ecosystem.

Considering that the demand for annotated and labelled data sets is likely to increase, a comprehensive understanding of the labour, business models and opportunities in this sector is critical. It would inform investment in skilling and infrastructure, approaches to encoding worker protections, as well as the policy interventions required to power the advancement of **responsible AI**.

As AI grows more sophisticated, labelling tasks have evolved in their complexity, requiring that workers possess expertise and skill to annotate specialised data such as medical data, and use more complex tools for annotation. Moreover, there has been a shift in terms of the demography of workers from a moderate-income US-based workforce to workers in the Global South in countries like India (Ross et al., 2010; Graham, 2018; Murgia, 2019; Simon, 2019).

Recent years have witnessed emergence of a private industry around data labelling—companies like iMerit, Taskmonk, Tika Data, Samasource and ClassifyIt offer data labelling services, but their models for sourcing work and labelling are different from traditional crowdwork platforms. Diverging from the often contractual and informal nature of labour relations in legacy platforms, our findings show that these entities employ labellers full-time and invest in training and worker well-being.

This report: an exploration of the dynamics of data labelling in India

Given this dynamic reality, this report seeks to explore the practices that underpin the AI data labelling industry in India through the lens of private entities and likely future evolution. Uncovering the evolving landscape of AI data labelling gives us an opportunity to re-envision and structure this labour to be just and equitable, while generating employment for a diverse cross section of India's population. At the same time, it also complicates the narratives around automation which primarily focus on job loss. By throwing light on the emerging geographies of work—much of this new annotation work is digital, and concentrated in smaller towns and rural areas—it also interrogates the narratives around urbanisation and migration. Equally, this report is timely as India grapples with growing post-pandemic unemployment.

Our approach has been to speak to industry leaders/startups to understand some of the emerging practices and mechanisms of formal and informal governance. Further, on the basis of this, we offer a proposal for a wider set of policy recommendations with a broader commitment to development of a more ethical AI supply chain.

Research was conducted keeping in mind the recent changes in labour codes by the Indian Parliament. While these codes recognise gig and platform workers, they have not classified them under the traditional employer-employee relationship, thereby placing little to no obligation on the platforms to provide them with benefits (Surie, 2020). As many labelling businesses still depend on crowdsourced labour, these laws will likely impact the industry and labour relations.

Our recommendations consider how labelling can offer employment opportunities while upholding worker rights and enhancing their agency and well-being. We envision our future research building on the findings of this report by surfacing worker voices and experiences on these platforms.

The report is structured as follows: after a brief description of the method and the extant literature, it summarises the findings from the primary research. This has three components: a discussion of the business models, description of the data labelling and annotation work, and an exploration of the evolution of data labelling work and possible future trajectories. Lastly, the report highlights potential policy interventions for promoting employment through this sector and protecting worker rights.

This research has been conducted in collaboration with Professor Alex Taylor from City, University of London, and with funding from the Global Challenges Research Fund.

2

Methodology

The objective was to explore the state of the labelling or annotation industry with its varying business models and prevailing practices. Desk research was carried out along with semi-structured interviews with founders of data labelling companies.

This approach has enabled, in this rapidly evolving sector, a deeper understanding of the realities of data labelling and its potential trajectories, and provided insight into the nuances of how platforms are structured. This inquiry could also inform policy approaches to support the twin goals of economic empowerment and worker well-being.

2.1 Sampling & Selection of Interviewees

The findings of this report are largely based on in-depth, semi-structured interviews conducted with eight leaders of startups in India.

The sampling process involved identifying a set of 24 startups focused on data labelling or annotation which were either located or operating in India. Accounting for variation in industry, publicly-available information, business model and size, 15 were contacted. From this shortlist, eight startups were interviewed. Of these, six provide data labelling/annotation as a core service/focus, while the other two carry out labelling to build their respective products.

An initial roundtable discussion with three startups (iMerit, Playment and Taskmonk) was held to establish context around data labelling processes and to shed light on varying models employed. It was conducted in collaboration with Dr.

Alex Taylor on June 16, 2020. The roundtable discussion focused on understanding the processes of data labelling, including the challenges faced by these startups and explored the future of work in data annotation.

2.2 Interview Process & Analysis

The interviews were structured to understand: the vision of these startups; challenges in maintaining quality; the shifting nature of data labelling work; recruitment and training of workers, worker demographics; and the future of annotation work.

Accounting for travel restrictions and safety requirements in view of the Covid-19 pandemic, interviews were conducted remotely. Once complete, interview transcripts were coded to identify themes. The initial set of findings assisted in deriving a set of hypotheses, which further informed questions through an iterative process for the second round of interviews with a select set of companies. A final analysis based on the roundtable and interviews surfaced ten themes which outline types of business models, practices, and the potential trajectories of data labelling work.

2.3 Limitations

Due to Covid-19 constraints that limited access, we did not interview data labellers and a wider set of startups, a gap we hope to address in future research. We hope this research starts a conversation to understand more about the business models and the lived experiences of workers.

3

Literature Review

Technology and applications based on machine learning (ML) models have become ubiquitous in our lives—there is a range of examples from facial recognition cameras to document classifiers in our accounts. For applications based on ML to work, they must be trained on high quality labelled data sets which are labelled according to their "target concepts". For example, faces labelled in images or emails labelled as spam or not spam (Chang et al, 2017). Over the years, a common approach to carrying out data labelling tasks has been to rely on crowdwork platforms which engage a "geographically distributed workforce to complete complex tasks on demand and at scale" (Kittur et al., 2013: 1). These platforms offer a cheaper way of labelling datasets as the workers on these platforms are a large pool of non-experts who are given small sums of money compared to trained annotators demanding higher wages (Snow et al., 2008; Novotney & Callison-Burch, 2010).

The extant literature on data labelling can be divided into two strands. The first strand focuses on crowdwork platforms and their impact on workers. The second concentrates on gig work in general, and data labelling work is often portrayed as being part of the emerging gig economy (Gray and Suri, 2019; Anwar and Graham, 2020). We discuss each in turn.

3.1 Lived Experiences: Understanding Crowdwork Platforms

The first strand of literature focuses on crowdsourced labelling. On these platforms, work is carried out online by labellers. Known as crowdworkers, they are paid by their requesters via the platform (Kittur et al., 2013). This strand of literature looks

at the lived experiences of these workers on platforms like Amazon Mechanical Turk (AMT), Upwork, Appen, etc., and explores ways to enhance worker protection and rights.

While primarily imagined for one-on-one requests, crowdwork platforms present companies with a cheaper alternative for getting labelled data as the work is completed efficiently, and there are no associated costs of hiring an employee (Gray and Suri, 2019). But, as Gray and Suri (2019) show, workers are often underpaid, and do not get other forms of worker protection like minimum wages or paid leave. Their work highlights the importance of making work like data labelling (which the authors call "ghost work") more visible.

Along similar lines, Zannotto (2019) has drawn attention to the importance of human data labellers in development of AI systems. He argues for a responsible Artificial Intelligence or a Human-in-the-loop Artificial Intelligence (HitAI) that gives these workers or "knowledge producers" rightful credit (Zannotto, 2019: 244). For this, he proposes either convincing or forcing AI companies to give back part of the revenues to these "knowledge producers" through legal and technological mechanisms (Zannotto, 2019: 247).

While a large portion of literature focuses on the lack of visibility of data labellers and the negative impact of crowdwork on data labellers, there have also been a few studies which explore platform design to make crowdwork more productive and fulfilling for the workers. For instance, Kittur et al. (2013) suggest the possibility of designing crowdwork platforms to create career ladders for crowdworkers by recognising their ability to take on more tasks, and rewarding expertise through incentives such as permanent roles.

While this literature is relevant in understanding some of the challenges of crowdsourced labelling work, the reality is that data labelling is moving towards different business models. These models rely less on gig work. Accordingly, the section that follows explores the extant literature on gig work.

3.2 Gig Work

Another strand of literature looks at platform work and examines empirical realities of gig workers. This literature around platform gig work is helpful in understanding the dynamics and aspirations of a workforce whose work is allocated online via an algorithm, wherein the labour process is controlled remotely (Gandini, 2019).

This literature, like the previous strand, also explores the lived experiences of workers. Anwar and Graham (2020) in their research on gig workers in African countries examine the difficulties faced by gig workers precisely because of algorithmic control. Gupta and Natarajan (2019) also describe the contestations that underline technologically mediated work, where, despite the employment opportunities, there is an absence of benefits, stable pay, and ultimately, a reinforcement of class and gender hierarchies. They also explore the ways in which workers have exercised agency. Resistance occurs both within the workplace—where workers negotiate better wages by examining payment history—and without wherein workers exercise resistance using local and social media networks such as gig workers' Facebook groups.

While platform gig work offers insight into thinking about lived experiences in tech-mediated work, it does not necessarily accommodate the possibility of full-time work. It draws attention to precarity of the work; but consequently, does not problematise the experience of tech-mediated, but full-time, skilled work. Moreover, this literature focuses on work mediated online but carried out offline like delivering food or driving a cab (Prassl, 2018).

3.3 AI Data Labelling Work: Uncovering the Puzzle

Together, these strands offer insight into the invisibility of labelling work and the lived realities of tech-mediated work. However, data labelling is uniquely at the intersection of these two challenges—it possibly accommodates a range of employment arrangements, including crowd and gig work. Additionally, this work can employ non-English speaking populations living in small towns and rural areas (Gupta et al., 2012; Chopra et al., 2019). However, while potentially critical to the future economy, this work and sector are invisible. As India grapples with a spurt in unemployment post the Covid-19 pandemic, visibilising this work is critical to better protecting workers and generating employment.

4

Findings from Primary Research

The wave of digitisation and demand for labelled data sets coupled with the limitations of traditional crowdwork platforms have shaped a diverse annotation and labelling industry. Recognising this as an evolving space, the first section of this report seeks to outline this spectrum of business models. The following section examines data labelling work from the perspective of the startups, considering questions of worker well-being, participation and empowerment. The third section looks at the evolution of the data labelling industry in India.

4.1 Crowdwork to In-house work: Emergent Business Models in Data labelling

Crowdwork/Platform as a Service (PaaS)

Contemporary labelling and annotation startups have largely emerged in response to the poor quality, accuracy, lack of domain-specificity and security associated with legacy crowdwork platforms like Amazon Mechanical Turk, Upwork and Appen. Characterised by remote, contracted workforces who carry out 'micro-tasks', these platforms engage in minimal worker training, providing little or no benefits to labellers. They also grant clients with insufficient control over the annotation process, and therefore over the quality of the performed task. Within this structure, workers are atomised, and their performance and pay are typically governed by algorithms, leaving few opportunities for negotiation and recourse. The relationship between the platform and the worker in this model means these companies are not liable to provide any worker protection or abide by standard labour regulations.

While crowdwork platforms offer enterprises low-cost labelling at scale, most startups we interviewed argued that these benefits seem to extend only to tasks that require low context and skills as well as for data sets that are not sensitive. As a result, many small to medium businesses building their own Machine Learning models either rely on in-house teams or outsource labelling work to third-party startups who also have the ability to manage the labelling process from end to end.

Software as a Service (SaaS)

The cost of labelling in-house can often be prohibitive for small businesses who may have domain-specific requirements that necessitate technical expertise or specific tools. Startups like Taskmonk, operating with a Software as a Service model, enable their clients with the interface and tools for their teams to label in-house or support them in onboarding data labelling partners, often Business Process Outsourcing (BPO) or impact sourcing firms. This managed services model grants greater control over the annotation processes where clients can embed AI to speed up human annotation and optimise the labelling budget to ensure better-quality output. In this model, the platform does not directly engage with workers, but instead interfaces and creates feedback loops between the BPO or labelling partner.

Hybrid

Recognising that businesses often require a combination of low context and skill labelling suited to crowdwork and more specific high-skill annotation, hybrid models engage workers both as contractors and full-time employees. The latter are often recruited as they possess particular skill sets that may be required to label forms of complex medical or language data. Workers are typically assessed for this knowledge and also provided training at the time of onboarding. In some cases, this structure also allows for contract workers to be employed full-time. Broadly, the hybrid model enables the scalability of labelling tasks possible with crowdwork platforms; however, it also ensures better data security for clients and quality control mechanisms, often unavailable with legacy platforms.

In-house

Most distinct from crowdwork platforms, startups that align with this model tend to prioritise the provision of domain-specific annotation services and consequently invest heavily in recruiting and training their own workforces. Labour is structured as a full-time opportunity with benefits and employees often work on-premise instead of remotely. Companies that adopt this model often also place importance on building inclusive and supportive work cultures which tend to reflect in additional welfare programmes and initiatives. Such startups we interviewed argued that

clients who outsource their labelling work to them can count on greater security measures and are better placed to adapt to new and evolving demands that may arise.

Two other startups that were interviewed did not carry out annotation or labelling as a core part of their business but, rather, engaged with these processes in-house in order to build Machine Learning models or Natural Language Processing (NLP) products. These startups are among a growing set of businesses that have a specific set of annotation requirements which are difficult to outsource to third-party providers. For instance, to build reliable speech recognition engines, language data is often pre-labelled at the point of collection. Post-collection, further annotation and quality control necessitate a greater degree of manual intervention by workers who possess the necessary linguistic knowledge, ontology and understanding of formatting.

Figure 1: Spectrum of data labelling businesses

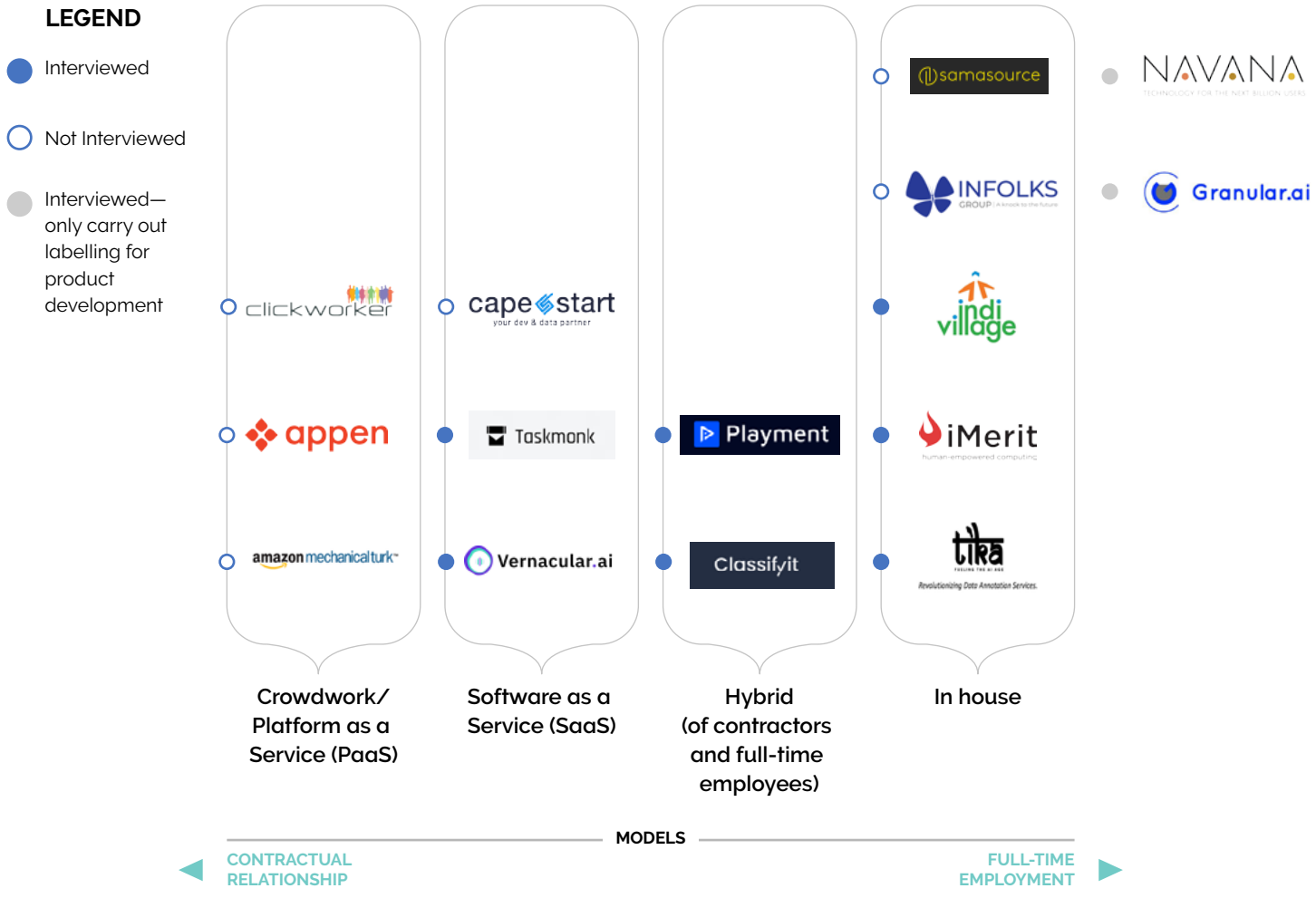


Figure 1 in the previous page identifies a range of businesses in the data labelling industry based on the models described above. It represents the labour relations of data labelling startups. From left to right of this spectrum, the relationship becomes less contractual and takes the form of employer-employee relations. On the right, the Hybrid model and In-house model recruit, train and skill workers, which lends greater control over the quality of labelling. This is in contrast with those on the left, the PaaS and SaaS models which enable quicker, low-cost and scalable annotation. The latter gives clients customisability and choice over onboarding vetted data labelling partners. These data labelling partners often include those that carry out labelling in-house.

4.2 Safety, inclusivity, and well-being: practices in the industry Data Labelling

In this section, we examine how businesses are structuring work, providing professional or technical development/training as well as securing their employees' rights and well-being. Without these considerations, this form of digital work, whether carried out remotely or in-house, may run the risk of mirroring or amplifying inequalities that exist offline, thereby further excluding and limiting workers' bargaining power. Based on our interviews, we highlight practices startups employ to surface worker voices, address needs and build more diverse and inclusive workplaces—lessons that are valuable to shaping discussions and policy on the future of work and workers.

Building a decentralised work culture for greater communication and collaboration

Crowdwork platforms like AMT have lacked communication channels between workers and requesters through which workers can share and receive feedback on the work they are doing (Gray and Suri, 2019). This has often resulted in unclear instructions affecting the quality of the labelled data sets (Kittur et al., 2013).

Findings from our interviews demonstrated the importance and value of building smaller teams with decentralised structures that give workers greater ownership over their labour, facilitate increased collaboration, and enable product improvement or innovation. At IndiVillage, labellers are encouraged to attend calls with clients to have visibility of and be connected with the outcome of their labour. Chirasmita Amin (IndiVillage) said that this exercise had led to greater motivation and empowerment of workers.

Jai Natarajan (iMerit) and Muzammil Hussain (Tika Data) described how small teams led by one manager make it easier to communicate across teams, assess progress and provide support more readily, at the operational level. For managers, this structure allows for greater oversight and points of interaction that allow challenges to be surfaced.

Further, these structures provide additional benefit for workers when coupled with formal or informal networks for worker communication and collaboration. For gig workers in Africa, Anwar and Graham (2020) found that labellers use social media to form networks to interact, seek guidance and even share tips for securing more contracts. In the case of Indian workers, Mawii and Aneja (2020) talk about the importance of online tools, forums and social media groups formed by workers which allow them to secure better information about their work, identify the right work opportunities and build solidarities with workers globally. This helps them in overcoming the alienating effects of crowdwork.

On similar lines, startups like ClassifyIt recognise and formalise these forums or chat groups which allows them to gain greater insight into platform and tool usability, allocate work opportunities and build solidarities with workers. While we cannot infer their actual impact on workers without deeper empirical investigation, they illuminate the ways in which grievance redressal mechanisms for workers can be conceptualised.

Decentralised organisational structures also help workers voice their grievances more quickly and effectively and receive feedback. At Tika Data, workers can directly raise concerns with the internal HR counsellor or during town hall gatherings. Rishabh Ladha (ClassifyIT) described support chat groups with the team leads where workers can discuss their concerns. Vernacular.ai has implemented a more formal mechanism, an Employee Satisfaction Council, composed of representatives of every team who discuss and frame company policies. Similarly, iMerit has launched employee focus groups and an Ethics Board which helps assess decisions that can impact employee well-being and welfare.

Providing opportunities for economic empowerment through training & skilling

Deviating from crowd work models, startups are increasingly investing in training and skilling workers to produce better-quality output. While some startups argue that investing in extensive training can be expensive, considering employee turnover, others like iMerit believe it contributes to worker loyalty, motivation and lends the organisation a competitive edge, enabling easier adaptation to more complex labelling demands from clients. Natarajan (iMerit) explained:

> "As algorithms get more capable, people have to become even more capable to stay ahead of them. The knowledge boundary keeps shifting and you have to work with the same people, train them in order to reap the benefits of that increasing expertise".

While training differs in approach, intensity and frequency, organisations like IndiVillage, ClassifyIt and Vernacular.ai follow a hybrid approach of both training and use of tools to assist their workers. iMerit adopts a dynamic training approach which encourages peer to peer learning and expert-led training. Though perhaps less formally structured, our research also uncovered that several startup leaders provide employees with professional development and career guidance, encouraging their employability in the data science industry.

Fostering safe and inclusive spaces to encourage diversity

The flexible and often remote nature of labelling work has meant that college students and stay-at-home mothers feature as common worker demographics within this industry. Ladha (ClassifyIt) said that the latter are typically highly skilled, trained and possess a greater motivation to work in order to supplement family income. Gray and Suri's research (2019) further reinforces how these forms of digital labour can provide marginalised groups with "digital literacy, a sense of identity, respect among family, and financial independence" (Gray and Suri, 2019: 125). However, without additional support—like paid family leave or affordable childcare—the burden of unpaid labour that often falls entirely on women makes it challenging for them to participate to the same degree as their male counterparts. This reality was particularly visible for women workers at IndiVillage during the initial Covid-19 lockdown. Factoring in childcare and other responsibilities, women worked longer hours at home and mentioned to leadership that they would rather work out of an office than remotely.

Recognising these challenges, IndiVillage provides workers with a creche facility and mental health channels for one-on-one counselling support. They also facilitate "lean-in" circles, enabling group discussions for women. On similar lines, Natarajan (iMerit) spoke about the importance of creating a 'safe and inclusive environment' which encourages women to grow and take on leadership roles in the organisation. iMerit has put this into practice by creating a women-only centre that - he asserted - workers defined as "a safe haven" in Metiabruz, Kolkata.

Aside from companies like iMerit or IndiVillage whose impact metrics necessitate gender diversity and women's empowerment, few others have specific policies in place to encourage leadership or enable participation. In order to prevent the reproduction of gender-based inequities, startups could consider additional measures to further support the hiring and meaningful inclusion of women by building safe spaces.

Embedding and prioritising initiatives that enhance worker well-being and agency

Data labelling work can often involve interacting with highly graphic or violent content. Such work can have a lasting impact on workers' mental health, taking an emotional and psychological toll (Gray and Suri, 2019; Roberts, 2019).

For work of this nature, startups vary in whether or not they take on these requests; many also provide workers with a choice regarding carrying out these tasks. For instance, Tika Data employs a large number of women and so based on their ethics guidelines they do not label pornographic content. Similarly at iMerit, labellers can decide whether they will annotate graphic material which can include images of violence or lifelike medical imagery, etc. This work is also guided by measures set up by an ethics committee in which workers also participate. Natarajan (iMerit) reflected upon the difficulties of dealing with violent data. He stated that there are two considerations that govern such work—whether the work makes the ecosystem better, and who would be able to do the work. Doing such work can cause "endless trauma" and make workers "get desensitised". iMerit governs this by granting workers agency over the decision to be on the project or not— employees can move to another project if the work causes discomfort.

Both companies offer counselling services but with Tika Data it only comes in the form of internal counselling whereas iMerit provides both internal and external, professional counselling.

4.3 Evolution of AI data labelling work and emerging trajectories

Owing to the evolving landscape of Artificial Intelligence, businesses once dependent on crowdsourcing platforms for data labelling and annotation are beginning to instead create new models or outsource these tasks to these third parties. Considering these changing conceptions of annotation work in the industry and the implication for labour relations, it is critical to chart future trajectories. Moreover, given that this industry presents significant sources of employment in India, it is important to assess how to shape these opportunities and structures in a way that grants agency and upholds the rights of both current and future workers.

The shift away from crowdwork

Our analysis found that companies are moving away from crowdwork platforms to in-house teams (Tika Data and iMerit) or to the Managed Services Model (Taskmonk). This is because labelling and annotation tasks are becoming more advanced, and require more skilled annotators that crowdwork platforms often

cannot provide. They also cannot guarantee the same degree of quality as trained and specialised in-house employees. For example, at iMerit (2020), in the case of annotating medical imagery, workers must possess specific skills that allow pattern recognition, and recall precise labels associated with healthcare ontologies. Training for these tasks is based on curricula developed by experienced medical professionals. This often starts with learning about simple annotation tools from 2D images and then graduating to multi-planar navigation in 3D imaging and on to 4D cine studies (iMerit, 2020).

According to Hussain (Tika Data), profit pools or value in data labelling seem to lie in handling proprietary data that requires safeguards and techniques to prevent re-identification, and illegitimate third-party data-sharing or breaches.

Moreover, the shifting knowledge boundary increases the costs of quality check in a crowdsourced model. For certain types of tasks to be profitable, an insourced model works better. While talking about how in-house employees are better than crowdsourced freelance workers, Natarajan (iMerit) pointed out that given the shifting knowledge boundary of AI data labelling, a firm needs the same set of people in order to have workers who are experts in the work which is only possible with in-house employees.

But there has not been a complete shift from crowdwork platforms as they are still a more cost-effective model, likely to be leveraged for simpler tasks that require lower specialisation or do not pose security concerns, like training data for autonomous vehicles. Additionally, companies are beginning to adopt measures like anonymisation, encryption and aggregation of data sets which have helped overcome security concerns over crowdsourcing platforms.

Increased interest in data labelling as a profession

We are also witnessing a change in the way data labelling is perceived—from being a form of 'invisible' labour (Gray and Suri, 2019), firms note a growing recognition of this labour. The demand for data labelling is such that Natarajan (iMerit) talked about the easy access they have to a larger pool of "specialists" from different professions like banking, retail and even medicine who are applying for data labelling jobs. Similarly, Natarajan argues that data labelling for women has become a profession which actually gives them greater recognition within their families. Particularly in the work-from-home context necessitated by Covid-19, the increased visibility of this labour evoked respect from family members who now understood the hard work their relatives did in their offices. The impact of this profession on women's autonomy and potential empowerment is also beginning to be documented and highlighted by startups like iMerit, IndiVillage and ClassifyIt.

Opportunities in small towns and rural areas

Research conducted by Chopra et al. (2019) and Gupta et al. (2012) has also demonstrated the employment potential of data labelling work in small towns and rural areas. We found in our interviews that this view was shared by startups that already envision labelling work being carried out by workers in Tier 3 and 4 towns and rural areas. Hussain (Tika Data) points out how people in rural areas might see it as a lucrative opportunity, given that

> "a Rs 10,000 salary, a health insurance and an air-conditioned office, and working on computers might be a better opportunity than doing manual labour in the fields".

IndiVillage has already set up a centre in a small village in Andhra Pradesh that employs three hundred workers in data labelling and annotation services.

Moreover, with the availability of distributed digital infrastructure (increase in internet connectivity and electricity), companies see this as a more cost-effective model due to the low cost of living outside metropolises and subsequently lower wages. However, according to Ladha (ClassifyIt), more specialised annotators will likely still be recruited from larger cities.

The human-in-the-loop is here to stay

As annotation tools, platforms and software advance in precision, trajectories over where the 'human-in-the-loop' will be located in labelling processes diverge. We identify two ways in which the AI data labelling industry may change.

Firs, increasing automation will not remove the human-in-the-loop, but will change the role the human will play in this supply chain. Automation's relevance lies in eliminating instances of human error and potential bias. This is giving rise to semi-automatic labelling, where annotators are given a head start and guidance in more accurately classifying data (Lee, 2020). Even the process of reaching consensus, a necessity amongst data labellers, is being automated using software like Revolt to generate better-quality labelled data sets (Chang et al., 2017).

Amin (IndiVillage) pointed out that workers need to be trained in order to develop skills enabling them to carry out highly skilled data labelling work, since repeatable and easy to carry out labelling tasks will become more automated. Both our interviewees and the literature around AI data labelling (Ruckenstein and Turunen, 2019) also point out that since humans are better suited to subjective understanding and dealing with ambiguity, they can provide an extra layer of quality check.

Secondly, human labellers are still critical for handling more complex cases. For instance, as pointed out by Jai Nanavati (Navana Tech), the labelling of language data presents a range of challenges like code-mixing, dialects and accents which only a human expert well-versed in the language can tackle.

Besides, as our interviewees argue, AI still hasn't developed the emotional intelligence which humans possess. As Sourabh Gupta (Vernacular.ai) points out,

> "We are working at the cutting-edge of AI but still the emotions and human behaviour, complexity, etc, I think AI is far from doing (replicating) it."

5

Balancing employment generation with worker protection: policy pathways

This section explores how opportunities for employment generation can be balanced with worker protection. We propose policy pathways for employment generation that advance the agenda of balanced development, which is inclusive of rural areas and women.

5.1 Considerations in policy-making:

Employment Generation

The data labelling industry in India, which is projected to reach $1.2 billion by 2023 (Bhatia, 2019), has the potential to generate employment as increasing automation and growth in Artificial Intelligence raise the need for more labelled data sets. This is especially important given that India has been facing growing unemployment—for the first time in its independent history the total number of employed people declined between 2011-12 and 2017-18 (Mehrotra and Parida, 2019). The pandemic has worsened this situation and 21 million salaried jobs were lost in April-August 2020 (Misra, 2020).

In parallel, the pandemic has also resulted in a push for digitisation in health, travel and hospitality which has consequently increased the demand for labelled data sets. This positions the AI labelling and annotation industry as a significant source of employment. Realising this opportunity, however, requires the industry to take note of the implications of the pandemic for both startups and workers and the resiliencies that have formed in response.

Initially, the industry was adversely impacted by social distancing and lack of infrastructure to enable productivity in a 'work-from-home' setting, but startups largely responded swiftly to these restrictions and even thrived in the midst of these challenges. Companies like iMerit have shown that they can train and upskill workers from marginalised communities and ensure they are equipped and have access to resources to carry out their work efficiently. These learnings are useful in a post-pandemic scenario where work-from-home opportunities are likely to be increasingly available and attractive for employers (Yadav, 2020).

Balanced Development

Promoting the AI data labelling industry also offers an opportunity to push for balanced development, both in terms of generating more job opportunities in small towns and villages as well as increasing the labour force participation of women.

Employment away from the metropolises—in small towns and rural areas

iMerit, IndiVillage and Infolks are employing people in smaller cities like Ranchi, Visakhapatnam, Yemmiganur and Shillong. iMerit employs around 200 workers from a small village, Kumaramputhur, in Kerala (Murali, 2019). Tika Data is pushing to build data labelling centres in rural areas by investing in setting up power and internet infrastructure. This allows companies to leverage the labour cost advantage to offer better and cheaper services. Startups like IndiVillage aspire to bridge the rural-urban gap by employing people from rural Andhra Pradesh.

Moreover, there is a specific opportunity in the context of small towns and rural areas in the field of Natural Language Processing (NLP) and Automatic Speech Recognition (ASR). This is because many people know both English and their local language, and are familiar with operating computers or mobile phones, which can allow them to create the much-needed training data sets for local languages (Kalyanakrishnan et al., 2018; Gupta et al., 2012; Chopra et al., 2019). Moreover, the demand for local languages has increased tremendously as more people from small cities and towns are going digital, requiring startups that can help in producing AI solutions which can cater to this demand (Ahskar, 2020). This has necessitated the emergence of service provision (e.g. Vernacular.ai and Reverie) that caters to the government and private sector language needs, and employs a significant number of people.

Gender diversity & inclusion in annotation work

While data labelling work has traditionally been seen as part-time gig work (Gray and Suri, 2019), we found that full-time women workers play a central role in the AI data labelling industry. Several of our interviewees indicated a preference for women over male employees—and women have come to occupy leadership roles in some of the companies. Therefore, the industry offers the opportunity to bring women into the labour force.

5.2 Potential Policy Pathways

As the industry is in a nascent stage, there is scope to imagine this work in a way that keeps worker well-being in the forefront. There is a need to make this work more visible, and give workers greater agency so that data labelling can be considered a lucrative career in which an individual can learn new skills and have scope for professional growth. Policy interventions need to balance these considerations, while encouraging market actors.

While the recently passed Code Bill on Wages (2019) and the Code on Social Security (2020) mention platform work (Surie, 2020; Sharma, 2020), they do not account for the emergence of the AI data labelling industry. Some recommendations:

Encode worker protections

The government should provide workers protections such as the right to select and reject work if it endangers physical and mental health, along with the provision for mental health counselling in the Code Bill on Social Security (2020).

This is especially critical in data labelling work as, unlike traditional factory work and other types of gig work carried out offline, labelling work is both managed and carried out online. This makes health and safety considerations differ from other forms of labour. For instance, data labellers can sometimes spend hours watching sensitive content in order to remove pornographic content or flag videos containing hate speech or graphic violence. This can have severe consequences on their mental health (Reese and Heath, 2016). Given this reality, the ability to select and reject work along with the provision of accessible mental health services is instrumental in ensuring the well-being of workers. Some startups in the industry (iMerit) already offer counselling services and give employees the choice of whether or not to take on these forms of work.

The new code which aims to secure bargaining power for workers is also a significant pathway to supporting worker well-being and rights. This can be done by mandating avenues for workers to voice their opinions regarding the difficulties they face in their work, the right to refuse work and collaborative forums which decide company policy including future projects.

Build worker capacity

Alongside regulating for worker protection, it is critical to focus on capacity building and skilling policies to strengthen the workforce. The National Skill Development Centre (NSDC) can explore the establishment of hubs in villages to train the rural workforce in operating computer and even mobile applications (Newlands and Lutz,

2020) for data labelling, especially NLP work. The government has already set up the Atal Tinkering Lab (ATL) AI Base Module and ATL AI-Step Up Module to bring in AI learning at school level. It has also created the National AI Portal of India which has resources for anyone who wants to learn the basics of big data and AI. Widening the reach of these programmes can help subsidise the training costs, and encourage new investments in this area.

Support to the industry

AI data labelling offers an opportunity to distribute employment geographically. Exploring government investment in requisite power and IT infrastructure in rural areas is a critical first step.

Our analysis indicates that in the short term, Covid-19 resulted in startups like Tika Data delaying plans to expand operations to set up centres in rural areas. Like other businesses, it has also forced startups to shift their operations from offices to homes. But due to limited infrastructure (electricity, internet, devices) and space in the homes of the data labellers, there are challenges in recreating an office environment. While startups are beginning to make headway in addressing this, further support could be provided by the government to secure these resources and capacity at scale.

In our interviews startups also recognised the value of government support, and pointed out the importance of making grants and subsidies accessible, and of tax relief for generating livelihood opportunities in rural areas.

6

Conclusion

Recognising the need to shape an equitable artificial intelligence supply chain, this report has aimed to shed light on an integral yet often less visible facet of this structure: data labelling. Following a literature review and synthesis of interviews with startup leaders, this report has identified emerging business models in the annotation industry, outlined key data labelling practices and projected how this space may evolve. These learnings have formed the basis of proposed policy interventions that highlight employment opportunities in this sector and suggest pathways to facilitate worker well-being, diversity in the workplace and rural development.

Insights in the report and emerging practices have been derived from conversations with founders and leaders of startups. This perspective enabled us to gain a deeper understanding of how the data labelling industry continues to unfold in India. Further research is critical to surface the lived realities and experiences of workers, and build on this work to uncover nuances associated with domain-specific annotation work.

While conversation and policy are beginning to evolve with respect to gig workers, we hope this report similarly sheds light on and fosters conversation around the labour undertaken by annotators. It aims to further serve as a foundation to acknowledge how this powers our AI-driven future and consider how these opportunities can be structured so as to be just and equitable, thereby shaking up narratives around automation and job loss.

7

References

Ahaskar, A. (2020, November 11). "Cos reach out in local lingo". *Mint*. https://www.livemint.com/companies/start-ups/cos-reach-out-in-local-lingo-11605057231014.html

Anwar, M. A., & Graham, M. (2020). Hidden transcripts of the gig economy: labour agency and the new art of resistance among African gig workers. Environment and Planning A: *Economy and Space*, 52(7), 1269–1291. https://doi.org/10.1177/0308518X19894584

Chakravarty, D. (2020, April 27). "Thinking about a career change during the coronavirus pandemic? Here are 8 things to consider". *The Economic Times*. https://economictimes.indiatimes.com/wealth/earn/thinking-about-a-career-change-during-the-coronavirus-pandemic-here-are-8-things-to-consider/articleshow/75371025.cms

Chang, J. C., Amershi, S., & Kamar, E. (2017). Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 2334–2346). Association for Computing Machinery. https://doi.org/10.1145/3025453.3026044

Chopra, M., Medhi Thies, I., Pal, J., Scott, C., Thies, W., & Seshadri, V. (2019). Exploring Crowdsourced Work in Low-Resource Settings. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300611

Cognilytica Research. (2020). *Data Preparation & Labeling for AI 2020* (Document ID: CGR-DLP20). https://www.cognilytica.com/2020/01/31/data-preparation-labeling-for-ai-2020/

Das, G. (2020, June 23). "In a season of layoffs, here's where jobs can be found". Mint. https://www.livemint.com/news/india/in-a-season-of-layoffs-here-s-where-jobs-can-be-found-11592915831636.html

Gupta, A., Thies, W., Cutrell, E., & Balakrishnan, R. (2012). MClerk: Enabling Mobile Crowdsourcing in Developing Regions. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1843–1852. https://doi.org/10.1145/2207676.2208320

Gupta, S., & Natarajan, S. (2020). *Futures of Workers*. Aapti Institute. https://www.aapti.in/blog/futures-of-workers

Graham, M. (2018, January 29). "The rise of the planetary labour market—and what it means for the future of work". NS Tech. https://tech.newstatesman.com/guest-opinion/planetary-labour-market

Gray, M. L., & Suri, S. (2019). *Ghost work: How to stop Silicon Valley from building a new global underclass.* Boston: Houghton Mifflin Harcourt.

iMerit. (2020). *The Challenge Of Med Ai Annotation*. [White paper]. iMerit. https://imerit.net/whitepapers/the-challenge-of-med-ai-annotation/?utm_campaign=brand&utm_medium=organic&utm_source=medai&utm_content=challengemediaiannotation

Joshi, S. (2019, September 9). "How artificial intelligence is creating jobs in India, not just stealing them"*Times of India*. https://timesofindia.indiatimes.com/india/how-artificial-intelligence-is-creating-jobs-in-india-not-just-stealing-them/articleshow/71030863.cms

Kalyanakrishnan, S., Panicker, R. A., Natarajan, S., & Rao, S. (2018). Opportunities and Challenges for Artificial Intelligence in India. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 164–170. https://doi.org/10.1145/3278721.3278738

Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., & Horton, J. (2013). The future of crowd work. *Proceedings of the 2013 conference on Computer supported cooperative work* (pp. 1301–1318). Association for Computing Machinery. https://doi.org/10.1145/2441776.2441923

Lee, I. (2020, April 9). "Data Labeling For Natural Language Processing". *TOPBOTS*. https://www.topbots.com/data-labeling-for-natural-language-processing/

Mawii, Z., & Aneja, U. (2020). *Gig Work on Digital Platforms: Online Support Tools and Forums for AMT Crowdworkers*. Tandem Research. http://tandemresearch.org/publications/gig-work-on-digital-platforms-online-support-tools-and-forums-for-amt-crowdworkers

Mehrotra, Santosh and Parida, Jajati K. (2019) *India's employment crisis: rising education levels and falling non-agricultural job growth*. Working Paper. Azim Premji University, Bengaluru. https://cse.azimpremjiuniversity.edu.in/wp-content/uploads/2019/10/Mehrotra_Parida_India_Employment_Crisis.pdf

Misra, U. (2020, September 13). "Explained: Taking stock of jobs lost, sectors affected, and possible ways forward". *The Indian Express*. https://indianexpress.com/article/explained/explained-taking-stock-of-jobs-lost-sectors-affected-and-possible-ways-forward-6589813/

Murali, A. (2019, March 21). "How India's data labellers are powering the global AI race". *FactorDaily*. https://archive.factordaily.com/indian-data-labellers-powering-the-global-ai-race/

Murgia, M. (2019, August 8). "AI's New Workforce: The Data-Labelling Industry Spreads Globally". *Financial Times*. https://medium.com/financial-times/ais-new-workforce-the-data-labelling-industry-spreads-globally-f472cb1bac09

Newlands, G., & Lutz, C. (2020). Crowdwork and the mobile underclass: Barriers to participation in India and the United States. *New Media & Society*. https://doi.org/10.1177/1461444820901847

NITI Aayog. (2018). *National Strategy for Artificial Intelligence*. Discussion paper. Retrieved from: https://niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf

Novotney, S., & Callison-Burch, C. (2010). Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. Human Language Technologies: *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 207–215. https://dl.acm.org/doi/10.5555/1857999.1858023

Prassl, J. (2018). *Humans as a service: The promise and perils of work in the gig economy*. Oxford: Oxford University Press.

Reese, H., & Heath, N. (n.d.). "Inside Amazon's clickworker platform: How half a million people are being paid pennies to train AI. Tech Republic". https://www.techrepublic.com/article/inside-amazons-clickworker-platform-how-half-a-million-people-are-training-ai-for-pennies-per-task/

Roberts, S. (2019). Behind the Screen: Content Moderation in the Shadows of Social Media. New Haven; London: Yale University Press.

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers? Shifting demographics in mechanical turk. In CHI '10 Extended Abstracts on Human Factors in Computing Systems (pp. 2863–2872). Association for Computing Machinery. https://doi.org/10.1145/1753846.1753873

Ruckenstein, M., & Turunen, L. L. M. (2020). Re-humanizing the platform: Content moderators and the logic of care. *New Media & Society*, 22(6), 1026–1042. https://doi.org/10.1177/1461444819875990

Sharma, L. (2020, November 2). "A secure future for platform workers". *The Hindu*. https://www.thehindu.com/opinion/op-ed/a-secure-future-for-platform-workers/article32998179.ece

Simon, N. (2019, July 25). "Africa: The Hidden Workforce Behind AI". *Mantra Labs*. https://www.mantralabsglobal.com/blog/ai-in-africa-artificial-intelligence-africa/

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—But is it good? Evaluating non-expert annotations for natural language tasks. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 254–263. https://dl.acm.org/doi/10.5555/1613715.1613751

Surie, A. (2020, October 8). "Gig work and its skewed terms". *The Hindu*. https://www.thehindu.com/opinion/op-ed/gig-work-and-its-skewed-terms/article32797547.ece

Thomas, A. (2020, June 4). "Amid Automation, Will We See More Data Labelling Jobs In India?" *Analytics India Magazine*. https://analyticsindiamag.com/will-we-see-more-data-labelling-jobs-in-india/

Yadav, N. (2020, July 30). *Post Covid-19 world: How AI can ensure a more inclusive digital economy*. ORF. Observer Research Foundation. https://www.orfonline.org/expert-speak/post-covid-19-world-ai-ensure-more-inclusive-digital-economy/

Zanzotto, F. M. (2019). Viewpoint: Human-in-the-loop Artificial Intelligence. *Journal of Artificial Intelligence Research*, 64, 243–252. https://doi.org/10.1613/jair.1.11345

8

Appendices

8.1 Startups Interviewed

1 ***iMerit***
Kolkata
Interviewee: Jai Natarajan

2 ***Taskmonk***
Bengaluru
Interviewee: Sampath Herga

3 ***ClassifyIt***
Delhi
Interviewee: Rishabh Ladha

4 ***Tika Data***
Bengaluru
Interviewee: Muzammil Hussain

5 ***Vernacular.ai***
Bengaluru
Interviewee: Sourabh Gupta

6 ***Granular.ai***
Boston
Interviewee: Siddharth Gupta

7 ***Navana Tech***
Bengaluru
Interviewee: Jai Nanavati

8 ***IndiVillage***
Kurnool
Interviewee: Chirasmita Amin

8.2 Interview Questions

Mission/Vision

a) Tell me about your company—what is your mission?
b) In what ways do you distinguish yourself from legacy platforms like MTurk?
c) How is work structured on your platform?

Building your platform/product

a) What was the focus or priority in building your platform? [e.g. building tools, providing aggregated services, integrating training, quality assurance pipelines, sector-specific tasks]
b) Why do you focus on these specific verticals?
c) What have the challenges been?

Shifting nature of work and increasing complexity of tasks

a) How has data labelling work evolved?
b) What are the new trends and demands in the AI data labelling industry?
c) Are there any changes that clients have asked for?
d) Are tasks more complex? How have you responded to this?

Quality

a) How do you maintain quality? What processes does this typically entail?
b) How is quality judged by clients?
c) What are the challenges/problems you've experienced in this process?

Training

a) How does training work on your platform?
b) How do you enable your workers to enhance their skill sets?
c) Are there opportunities for growth?

Culture

a) Do you encourage collaboration amongst workers?

Ethical & Equitable AI

a) What does 'Human-in-the-loop' AI look like to you?
b) In what ways do you integrate this into your workplace processes and culture?

contact@aapti.in