MARCH 7, 2025

In-house data annotation best practice principles developed by MTI

by Isabelle Amazon-Brown and The MERL Tech Initiative

Event recap: Humans in the machine – the impact of AI on workers

On February 6th, the NLP Community of Practice's Ethics & Governance Working Group convened to discuss the impact of AI development on the (human) workers involved in developing and refining Large Language Models (LLMs). We also wanted to highlight current thinking on what can be done to counteract the labour abuses currently baked into AI supply chains, emphasising that negative impacts can also occur at a smaller scale, for example when adapting LLMs in-house for 'positive' reasons such as improving relevance and safety. The recording and resources for this event can be found here.

Our panellists were:

- Soma Mitra-Behura, Senior Data Scientist at Girl Effect, who is spearheading the organization's integration of GenAI into its sexual health chatbot for South African teenagers, Big Sis.
- Mophat Okinyi, Chairperson of the Content Moderator's Union, founder and CEO of Techworker Community Africa, and recently named one of TIME's 100 most influential people in AI 2024, who talked about his first-hand experiences of AI labour, and shared his thoughts on how we can all support data workers.
- Maria Mukobi, data curator at Girl Effect, who has been involved in the data annotation process required to implement Retrieval Augmented Generation (RAG) within Big Sis' AI model.
- Oğuz Alyanak, postdoctoral researcher with the Fairwork team from the Oxford Internet Institute, who have developed an assessment framework and guidance for evaluating the labour conditions of data platforms, including LLMs.

> “The act of simplifying reality for a machine results in a great deal of complexity for the human.”

Our discussion kicked off with a brief overview of the role of humans in Large Language Model development, a topic highlighted by this in-depth expose as early as 2023 and which has now gained mainstream recognition. The ongoing role of human input is encapsulated by a comment from a source that noted wryly that "*ChatGPT seems so human because it was trained by an AI that was mimicking humans who were rating an AI that was mimicking humans who were pretending to be a better version of an AI that was trained on human writing.*"

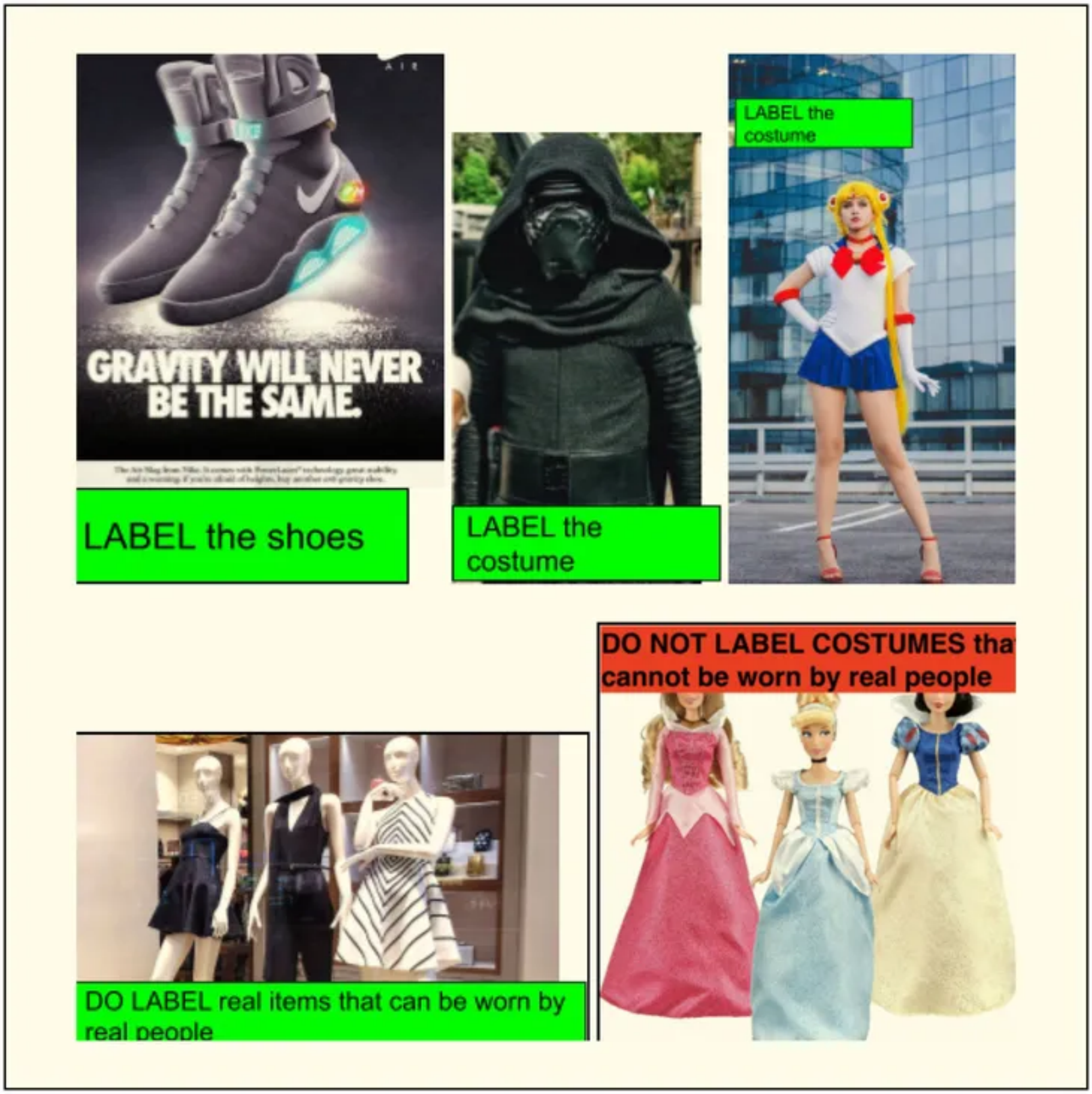
Soma provided a useful metaphor: a parent teaching a child to talk by helping them associate sights, sounds, smells or sensations with certain words is much like the heart of the data annotation process. In AI terms, each time an association is made between, say, the image of a dog and the word/sound 'dog', a new 'row' of data is created. But because, despite what we may *feel,* AI is not nearly as adept as humans at learning independently, the data it is fed needs to be formatted in hyper-specific ways. This is a second way in which data annotators perform an invisible function: *"AI has no methodology of exploring the world (…) so it's not just pointing out that a picture means 'mum' or 'dad', it's putting that association in a format that a computer will understand."*

These two steps alone represent one of the many 'layers' of data annotation that needs to take place to train an AI system. The additional stages are the equivalent of moving beyond individual word associations during our toddler years to understanding things like grammar and syntax as we develop. The role of data annotators as parents and teachers of AI, can't be overstated – and this is entirely at odds with the poor working conditions many who do this work experience.

The reality of data annotation for Global South Workers

Because of the significant human resources required to develop LLMs as powerful as ChatGPT and other commercial models, companies have yet again turned to the cheap workforces in LMICs – especially countries like Kenya, Uganda and India. These countries are specifically targeted because they often combine highly educated workforces with extremely high youth unemployment rates, as well as tax incentives and outdated or lax labour laws. Many young data workers are also lured by the promise of working on cutting edge technology, as Mophat highlighted: "I heard from my friends that AI is the future of technology, I wanted to be part of it – I was very curious about what AI is, and what it takes to train AI."

Unfortunately, the reality of the work was starkly at odds with the premise under which workers like Mophat were hired, by intermediaries such as Sama, themselves hired by organisations in the US recruiting anonymously on behalf of big commercial LLM companies. Firstly, Mophat's dream of feeling part of something big and exciting, and opportunities to upskill in a cutting-edge field, were hindered by the fact that workers are provided with no context for the work they are doing, often given seemingly nonsensical tasks with no insight into how the task is feeding into a bigger picture.


AIR
GRAVITY WILL NEVER BE THE SAME.
LABEL the shoes
LABEL the costume
LABEL the costume
DO NOT LABEL COSTUMES tha cannot be worn by real people
DO LABEL real items that can be worn by real people

An example of a relatively innocuous yet absurd-seeming data annotation task from The Verge

Additionally, the work itself was unreliable in terms of its frequency and duration, with projects starting and stopping at a moment's notice, and either poorly paid, or unpaid, based on spurious reasons. More egregiously, Mophat talked of the mental and emotional toll of data annotation, which involved labelling images of abuse or violence – causing nightmares and affective issues with a worker's children or family members. Even where organizations provide mental health support to workers, counsellors are not experienced enough in the impact of data-annotation induced trauma to provide adequate help.

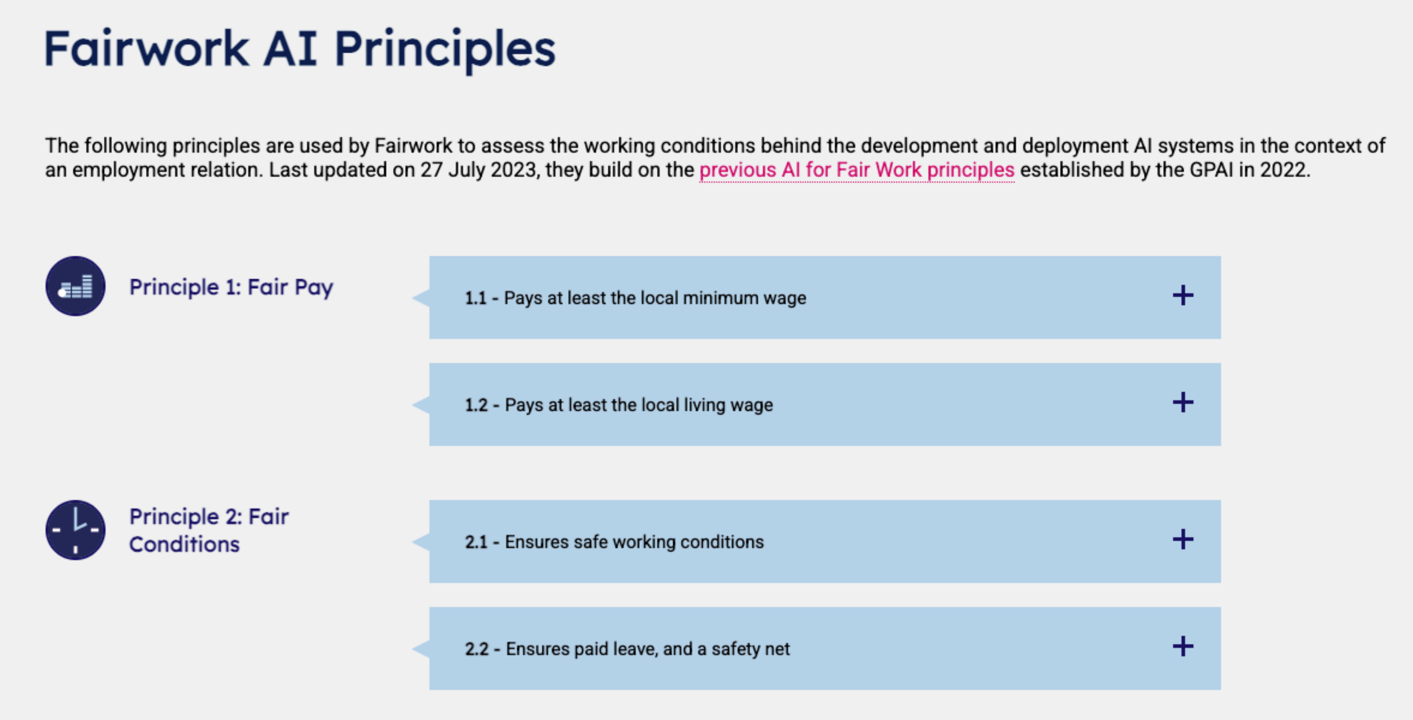
This impact is not unique to commercial LLM supply chains. Maria, who is involved in data annotation to make a commercial LLM more safe and relevant for young female South African users of the Big Sis chatbot, also spoke of the huge weight of responsibility she and her colleagues have faced in making decisions as a 'mother of AI'. Whereas Mophat spoke of feeling divorced from the impact of his work because of explicit efforts to keep workers in the dark, in-house data annotators can have an all-too strong connection to their work's potential impact.

For example, Maria spoke of the worry and guilt that could be caused by the knowledge that her decisions might end up directly shaping a GenAI powered response to, say, a young woman seeking help with the decision to terminate a pregnancy. Not only should this enormous responsibility be reflected in the status (and therefore pay) of data annotators, including at a smaller scale, but processes should also be put in place to prepare and support them to shoulder this burden on an ongoing basis.

Frameworks and tools to assess and respond to labour abuses

Our discussion moved on to focus on what could be done to address these problematic issues – either when choosing which LLMs to work with, or when conducting data annotation in-house.

Oğuz spoke of the Fairwork action-research project, which aims to highlight the best and worst examples of how technology supply chains impact workers, and offers core principles developed in conversation with scholars, policy makers and workers themselves. These principles are then used to publicly hold companies accountable. To date, Fairwork has evaluated digital labour conditions in 40 countries and produced over 700 platform ratings including Uber, or JustEat as well as AI intermediary companies like Sama.



Fairwork AI Principles

The following principles are used by Fairwork to assess the working conditions behind the development and deployment AI systems in the context of an employment relation. Last updated on 27 July 2023, they build on the previous AI for Fair Work principles established by the GPAI in 2022.

Principle 1: Fair Pay

1.1 - Pays at least the local minimum wage +

1.2 - Pays at least the local living wage +

Principle 2: Fair Conditions

2.1 - Ensures safe working conditions +

2.2 - Ensures paid leave, and a safety net +

2 of the 5 Fairwork AI principles

We also spoke of the delicate process of engaging with companies held up to scrutiny and doing justice to the demands and needs of workers they speak to. Oğuz shared a case study involving Sama, the same company Mophat had worked for. Fairworks presented Sama with a 'preliminary assessment' based on desk research, interviews with workers and conversations with Sama managers. Two points were awarded for each of Fairwork's 5 principles, based on issues such as unpaid overtime, short term contracts, dangerous levels of job strain and excessive workplace surveillance. Sama received a score of 0 out of 10.

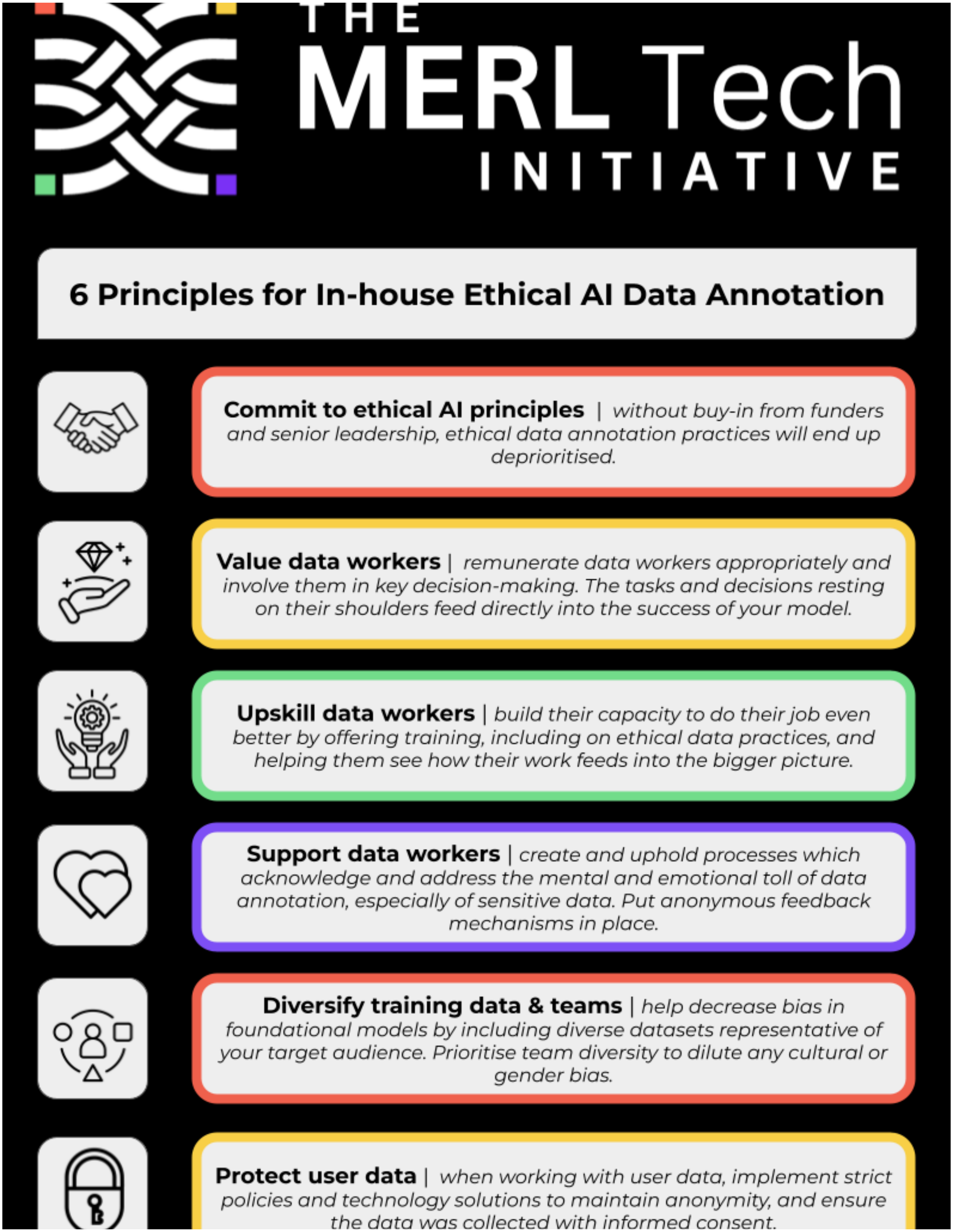
Sama showed willingness to engage and make changes, which did result in a 15% pay rise to the base salaries of almost 4,000 workers and improved access to psychological help – leading them to a score of 5 out of 10 (a 10/10 score is considered the 'bare minimum' in terms of rights-respecting workplace). Indeed, Mophat earlier mentioned that many Sama workers felt that the changes introduced by Sama were insufficient to merit a 5 out of 10. Oğuz agreed with this and used it as an example to illustrate the importance of ongoing, repeated assessments on AI supply chains.

Ultimately, Oğuz stressed Fairwork's plan to move from evaluating intermediaries, including recruitment companies that handle the advertising and allocation of digital work like data annotation or content moderation, to holding the 'lead firms' accountable too. These large companies often hide their involvement via multiple layers of outsourcing. They, however, are the ones who ultimately define the working conditions through their budget stipulations, timeframes, limitations on workers' rights via NDAs, and lack of upskilling opportunities.

Calls to action to improve data annotation conditions

Fairwork's global work will inevitably take time to move the needle, as it involves holding hugely powerful firms and complex systems accountable in a meaningful and enduring way. But closer to home, we may have more opportunities to improve working conditions for data annotators, starting with providing mental health support as standard. There is a tendency when working with new technologies to take a blank-slate approach, when actually what's needed is a doubling-down on existing best practices when it comes to workers' rights.

One simple step we can take is also to prioritise opportunities for education, upskilling, and certification. Mophat also stressed how important it is to educate workers not only about their rights but about the existence and benefits of organizing via trade unions such as Tech Worker Community Africa. Data workers need to be made aware of their own value in terms of the crucial role they are playing in developing AI models which ultimately enrich a few billionaires in the Global North and be given recognition for the complex skills they have developed (from scratch!). Mophat mentioned how frequently data workers are left 'stranded' with nothing to show for their unique experiences.



THE MERL Tech INITIATIVE

6 Principles for In-house Ethical AI Data Annotation

Commit to ethical AI principles | *without buy-in from funders and senior leadership, ethical data annotation practices will end up deprioritised.*

Value data workers | *remunerate data workers appropriately and involve them in key decision-making. The tasks and decisions resting on their shoulders feed directly into the success of your model.*

Upskill data workers | *build their capacity to do their job even better by offering training, including on ethical data practices, and helping them see how their work feeds into the bigger picture.*

Support data workers | *create and uphold processes which acknowledge and address the mental and emotional toll of data annotation, especially of sensitive data. Put anonymous feedback mechanisms in place.*

Diversify training data & teams | *help decrease bias in foundational models by including diverse datasets representative of your target audience. Prioritise team diversity to dilute any cultural or gender bias.*

Protect user data | *when working with user data, implement strict policies and technology solutions to maintain anonymity, and ensure the data was collected with informed consent.*

In-house data annotation best practice principles developed by MTI

These final calls to action should be heard by those of us working directly with young people who may themselves be tempted by data annotation roles in their respective countries. It is becoming clearer and clearer that anyone involved in using an LLM to power their digital interventions has a responsibility to simultaneously educate their audience members on the risks associated with the wider phenomenon of AI – whether that means using it or participating in some way in its development.

YOU MIGHT ALSO LIKE

If AI solves problems as well as creates them, can one counteract the other?: Honest Discussions at the Intersection of AI and the SDGs

by Bárbara Paes

The Humanitarian AI Countdown: How do humanitarian organizations operationalize Responsible AI with Shivaang Sharma (October 15th 2025 at 3pm BST/10am ET/ 5pm EAT)

by Bárbara Paes

Event Recap – Mapping connections, uncovering complexity: real-world applications of knowledge graphs

by Bárbara Paes

Introducing our new event series – “4,3,2,1: The Humanitarian AI Countdown”

by Bárbara Paes

Privacy