

THERE IS

NO "AI"

WITHOUT

DATA WORK

MILAGROS MICELI

✉ m.miceli@tu-berlin.de

🌐 milamiceli.com

🐦 [@milagrosmiceli](https://twitter.com/milagrosmiceli)

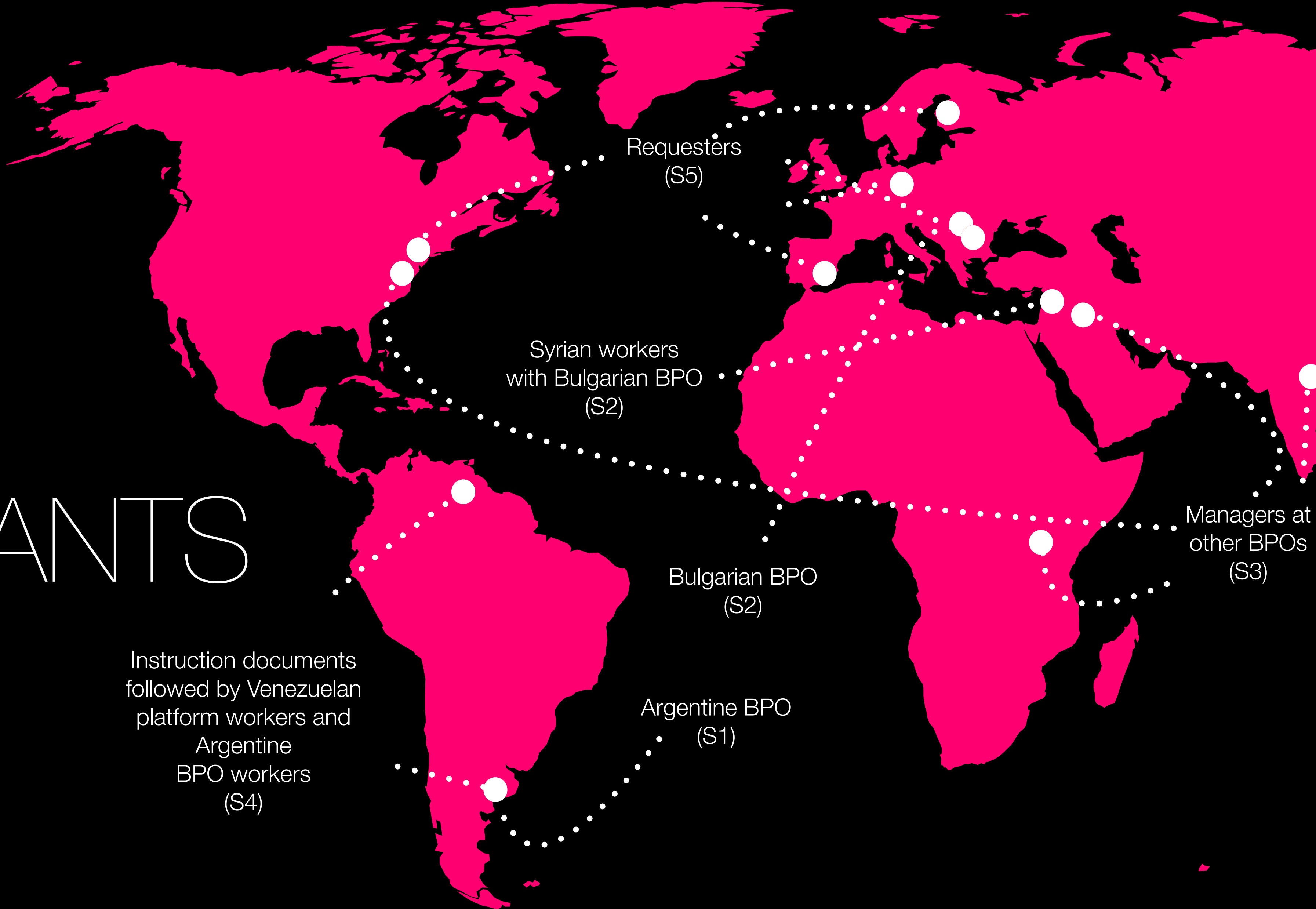
weizenbaum
institut





MIMI OUNOHA | "THE FUTURE IS HERE!"

SITES AND PARTICIPANTS



MILAGROS MICELI

 @milagrosmiceli

 milamiceli.com

DATA WORK TASKS



DATA GENERATION

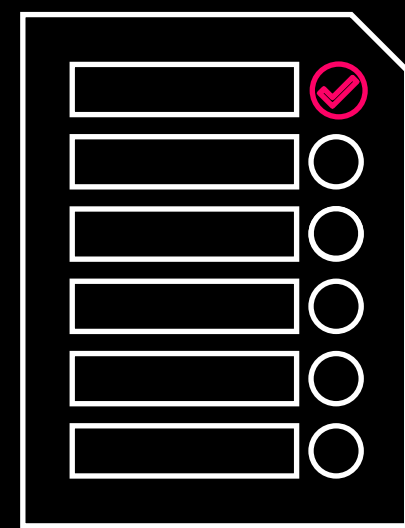
"You can earn \$2.5 by completing the task 'Do you wear glasses?' Upload a picture of a document with your prescription values"



ANNOTATION

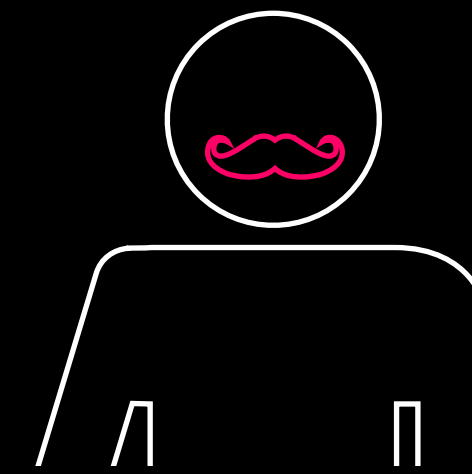
04/28

"Based on the text in each task, select one of these three options: Sexually Explicit, Suggestive, Non-Sexual"



VERIFICATION

"You'll be shown two lists of up to eight search suggestions each. Your task is to indicate which list suggestion is better"



IMPERSONATION

"As the assistant, the user will initiate the conversation [...] you need to use the facts to answer the user's questions."

Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic



”OpenAI SENT TENS OF THOUSANDS OF SNIPPETS OF TEXT TO AN OUTSOURCING FIRM IN KENYA, BEGINNING IN NOVEMBER 2021. MUCH OF THAT TEXT APPEARED TO HAVE BEEN PULLED FROM THE DARKEST RECESSES OF THE INTERNET. SOME OF IT DESCRIBED SITUATIONS IN GRAPHIC DETAIL LIKE CHILD SEXUAL ABUSE, BESTIALITY, MURDER, SUICIDE, TORTURE, SELF HARM, AND INCEST.”

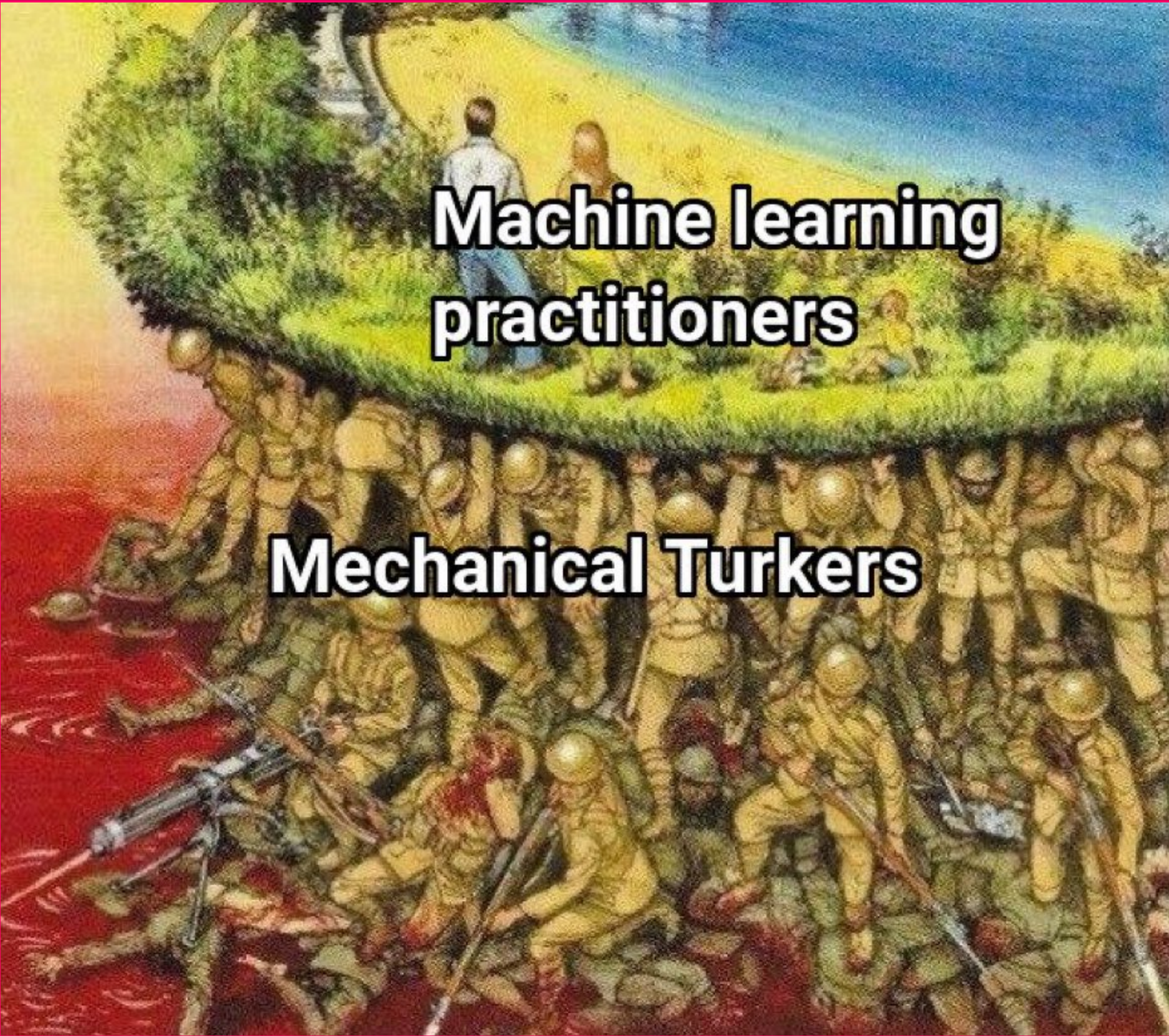
MILAGROS MICELI

 @milagrosmiceli

 milamiceli.com

✓ AI ETHICS

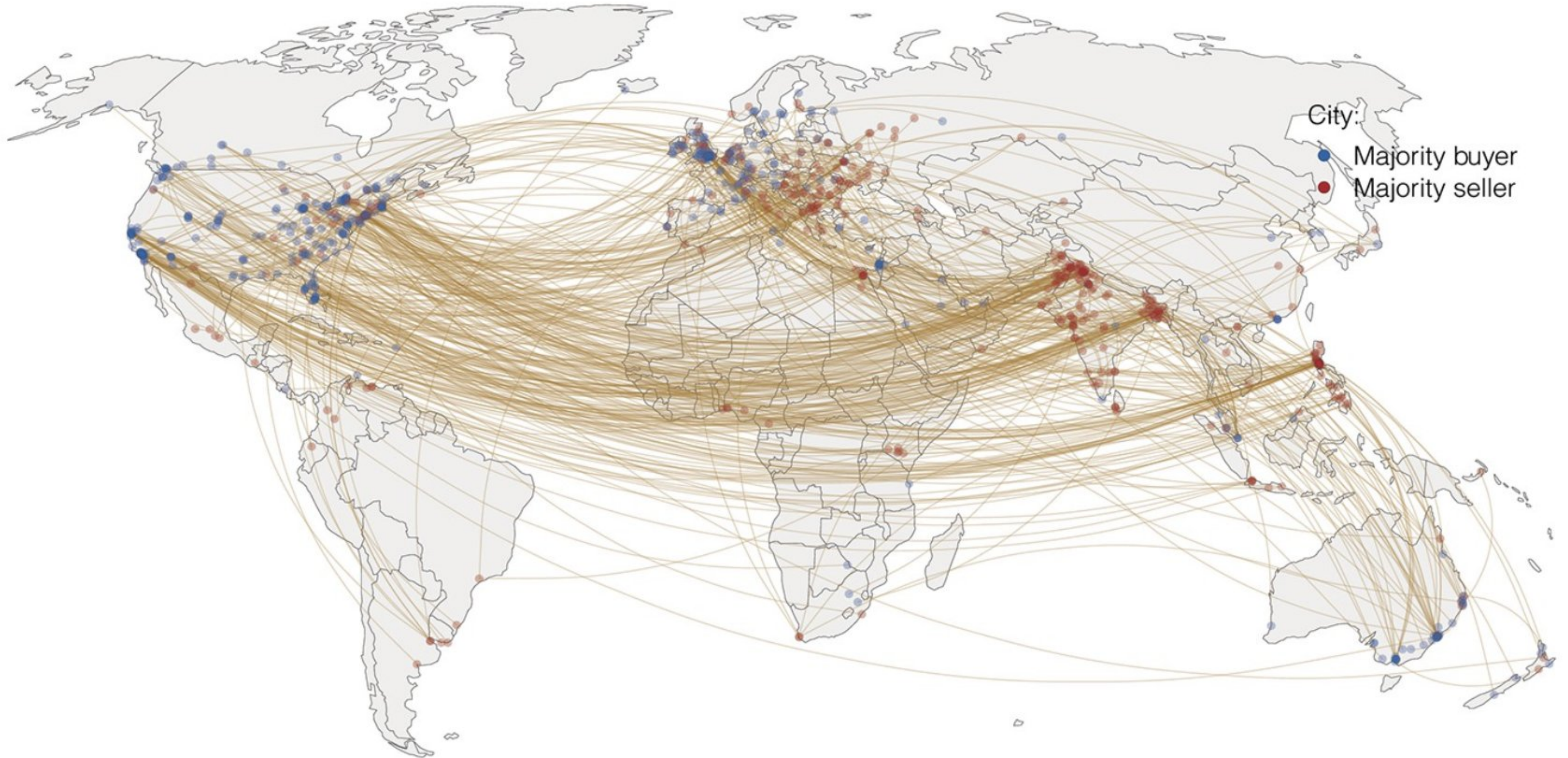
✓ DATA QUALITY



Machine learning practitioners

Mechanical Turkers

Flows in the global online labour market between cities with majority buyers (blue) and sellers (red) in 2020 (5% sample)



In this task, you will be identifying messages that contain **hate speech**.

Based on the text, you must select:

- **Hate Speech:** if the username contains hateful content
- **None:** if there is NO hateful or abusive language in the given [sic]

Definition:

Select **Hate Speech** if the text contains **any** of the following:

- Discrimination, disparagement, negativity, or violence against a person or group based on a **protected attribute**
- References to hate groups, or groups that attack people based on a **protected attribute**

TASK INSTRUCTIONS

Select:

- **Male:** if the boxed face is a male
- **Female:** if the boxed face is a female
- **Other:** if there is no face in the box

In this task you will be determining the **race** of the persons in the images.

You should select only **one** of the following categories:

- White
- African American
- Latinx or Hispanic
- Asian
- Indian
- Ambiguous

This is a **high paying job**, a special job, but to gain access to it and to keep access to it after passing the qualification test, we require patience and **VERY careful [sic] thought out and accurate responses.**

Otherwise, you will, unfortunately be banned from the job :(

Please keep your judgments consistent UNLESS you feel that there is some difference in the two that would result in a change of overall score.

Judges providing low quality responses will be banned and not paid.

✕ [Instructions](#)

1 Task Unit
\$ 0.55 per 1.000 task Unikats

🔄 Task Timer 00:00:39

Guess how old I am?

[Previous](#)

[Next](#)



Image Settings

Brightness [Reset](#)

Contrast [Reset](#)

Hotkeys [View Keyboard Shortcuts](#)

Categories

🔍 Type category namens to filter

- [Baby](#)
- [Toddler](#)
- [Pre - Teen](#)
- [Teenager](#)
- [Adult](#)
- [Middle - Age](#)
- [Senior](#)
- [Other](#)

EDIT PROJECT

Basic

Classes and attributes

Images and datasets

User and roles

Upload annotations

Automated labelling

MILAGROS MICELI

@milagrosmiceli

milamiceli.com

Users

Administrator ▾

User Roles

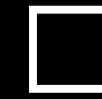
Project Owner
 Administrator
 Supervisor
 Labeller
 View Only

Administrate

Edit classes and attributes



Upload images



Export



Invite / Edit / Remove Users



Import



for
WISDOM ~~OF~~ CROWDS

MILAGROS MICELI

 @milagrosmiceli

 milamiceli.com

POWER ASYMMETRIES IN DATA
WORK FUNDAMENTALLY SHAPE
DATA AND SYSTEMS

MILAGROS MICELI

 @milagrosmiceli

 milamiceli.com

MICELI ET AL (2020), "BETWEEN SUBJECTIVITY AND IMPOSITION,
POWER DYNAMICS IN DATA ANNOTATION FOR COMPUTER VISION"

WORKER BIAS

Measuring Social Biases of Crowd Workers using Counterfactual Queries

Bayesian Bias Mitigation for Crowdsourcing

Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments

Crowd Prefers the Middle Path: A New IAA Metric for Crowdsourcing Reveals Turker Biases in Query Segmentation

MILAGROS MICELI

 @milagrosmiceli

 milamiceli.com

WHOSE BIAS?



*Ceci n'est pas une **bias***

MILAGROS MICELI

 @milagrosmiceli

 milamiceli.com

✓ AI ETHICS

MILAGROS MICELI

 @milagrosmiceli

 milamiceli.com

Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic



Is AI fast becoming a technology built on worker exploitation from Global South?

While working with crowd work platforms for datasets, it is essential to consider annotator subjectivity as it has the capability to make the data set of extremely high or low quality, which in turn affects the whole ML model.

ARTIFICIAL INTELLIGENCE

How the AI industry profits from catastrophe

As the demand for data labeling exploded, an economic catastrophe turned Venezuela into ground zero for a new model of labor exploitation.

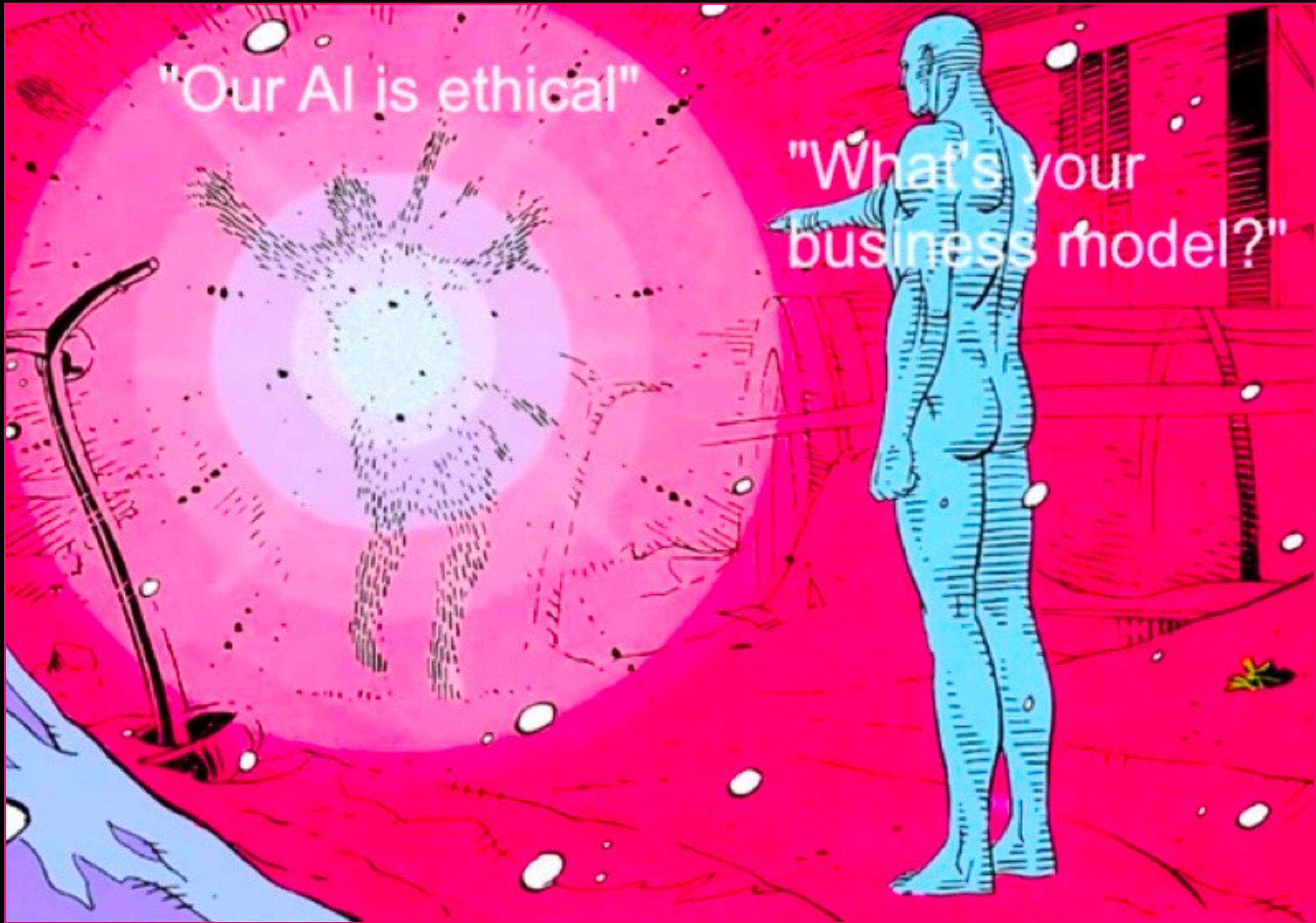
How low-paid workers in Madagascar power French tech's AI ambitions

Published: March 30, 2023 10.45am CEST

MILAGROS MICELI

 @milagrosmiceli

 milamiceli.com



@BLAIRASASERVICE

MILAGROS MICELI

 @milagrosmiceli

 milamiceli.com



“THOSE ARE MY
PRINCIPLES.
IF YOU DON'T LIKE
THEM...WELL, I HAVE
OTHERS.”

GROUCHO MARX

MILAGROS MICELI

 @milagrosmiceli

 milamiceli.com

✓ DATA QUALITY

MILAGROS MICELI

 @milagrosmiceli

 milamiceli.com

Data quality scales better than data size

Many of these projects are saving time by **training on small, highly curated datasets**. This suggests there is some flexibility in data scaling laws. The existence of such datasets follows from the line of thinking in **Data Doesn't Do What You Think**, and they are rapidly becoming the standard way to do training outside Google. These datasets are built using synthetic methods (e.g. filtering the best responses from an existing model) and scavenging from other projects, neither of which is dominant at Google. **Fortunately, these high quality datasets are open source, so they are free to use.**

ROLE PLAY: GENERATING QUESTIONS | GROUP A

🕒 10 mins. 👤 breakout groups 💡 hands-on brainstorming

Obtaining and sharing information: **Labelers**

Card 1 (Light Blue):
Topic: Ethics
Question: What are the implications of this data?
Who holds this information?: Requester

Card 2 (Yellow):
Topic: Classes for labeling
Question: How do you define each class?
Who holds this information?: Humans in the Loop

Card 3 (Light Blue):
Topic: \$\$\$\$\$
Question: *How much will I earn?
Who holds this information?: Ranua/AA/Requester

Card 4 (Yellow):
Topic: Labelling type
Question: **What are they labelling
Who holds this information?: Requester

Card 5 (Yellow):
Topic: Edge Cases
Question: What type of edge cases are included/not??
Who holds this information?: Requester

Card 6 (Yellow):
Topic: Goal
Question: What's the goal of the project? What is this data used for?
Who holds this information?: Requester

Card 7 (Yellow):
Topic: Annotation Precision
Question: *How precise the annotation should be?
Who holds this information?: Requester + AA

Card 8 (Yellow):
Topic: Outcome
Question: What will the project will be used in the future?
Who holds this information?: Requester
Relevance: little ————— very

Card 9 (Yellow):
Topic: *Desire Output
Question: What kind of annotation strype? Polygon, bounding box, etc Connected with the tool
Who holds this information?: Action Data

Card 10 (Yellow):
Topic: Instruction question & uncertainty
Question: *Do we need to follow a specific person? (ID technique) * How do we handle if an object is occluded? * What if you can't see an object * Are there parts of the image to ignore?
Who holds this information?: Requester

EXPERTISE AND DATA PRODUCTION

MICELI ET AL (2022), "DOCUMENTING DATA PRODUCTION PROCESSES. A PARTICIPATORY APPROACH FOR DATA WORK"

Project description with initial training material

Space for comments

The screenshot shows a project page layout. On the left, there is a video player with a play button and a list of text items below it. On the right, there is a comment section with several entries. Each entry starts with a user icon (a box with a letter) and a text area. The first comment is by user 'V', followed by 'S', 'J', 'V', 'O', 'PM', and 'V'. The 'O' comment includes a star rating of three stars. The 'V' comment at the bottom also has a star rating of three stars. A large red circle highlights the comment section.

Changelog 3 new events since last login

The screenshot shows a changelog table with four entries. Each entry consists of a date, a text description, and a status icon. The dates are 01/03/2022, 01/01/2022, 12/23/2021, and 12/10/2021. The status icons are a circled 'X', a circled 'X', a circled 'X' with two exclamation marks, and a circled 'X'. A red circle highlights the entire changelog section.

Date	Description	Status
01/03/2022	[Redacted]	⊗
01/01/2022	[Redacted]	⊗
12/23/2021	[Redacted]	⊗!!
12/10/2021	[Redacted]	⊗

[Link to annotation tool](#)

[Other tools](#)

[Link to data repository](#)

DATA EXPERTS

“IT IS NOT MERELY A
COLLECTION OF FORMAL AND
MECHANICAL RULES, BUT A
SITUATED AND DISCRETIONARY
PROCESS REQUIRING DATA
ANALYSTS TO CONTINUOUSLY
STRADDLE THE COMPETING
DEMANDS OF FORMAL
ABSTRACTION AND EMPIRICAL
CONTINGENCY.”

DATA QUALITY

Running Out of Time: The Impact and Value of Flexibility in On-Demand Crowdwork

Ming Yin, Siddharth Suri, Mary L. Gray
Microsoft Research
Ming.Yin, suri, mlg@microsoft.com

Incentivizing High Quality Crowdwork

Chien-Ju Ho
University of California, Los Angeles
cjho@ucla.edu

Aleksandrs Slivkins
Microsoft Research
slivkins@microsoft.com

Siddharth Suri
Microsoft Research
suri@microsoft.com

Jennifer Wortman Vaughan
Microsoft Research
jenn@microsoft.com

Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research

Jennifer Wortman Vaughan
Microsoft Research
641 Avenue of the Americas, 7th Floor
New York, NY 10011

JENN@MICROSOFT.COM

Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation

Lora Aroyo, Chris Welty

Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work

MIKE SCHAEKERMANN, University of Waterloo, Canada
JOSLIN GOH, University of Waterloo, Canada
KATE LARSON, University of Waterloo, Canada
EDITH LAW, University of Waterloo, Canada

- DATA WORKERS ARE THE FUEL THAT POWERS AI TECHNOLOGIES.
- PRIOTIZATION OF QUANTITY INSTEAD OF QUALITY.
- THERE IS NO ETHICAL AI WITHOUT ETHICAL DATA WORK.
- EMPOWERING DATA WORKERS AND LEVERAGING THEIR EXPERTISE CAN HELP PRODUCE BETTER DATASETS AND SYSTEMS.

TAKEAWAYS

MILAGROS MICELI

 @milagrosmiceli

 milamiceli.com

THANK YOU!

MILAGROS MICELI

✉ m.miceli@tu-berlin.de

🌐 milamiceli.com

🐦 [@milagrosmiceli](https://twitter.com/milagrosmiceli)

weizenbaum
institut

