The Data-Production Dispositif: Power/Knowledge in Outsourced Data Work for Machine Learning

Julian Posada and **Milagros Miceli**

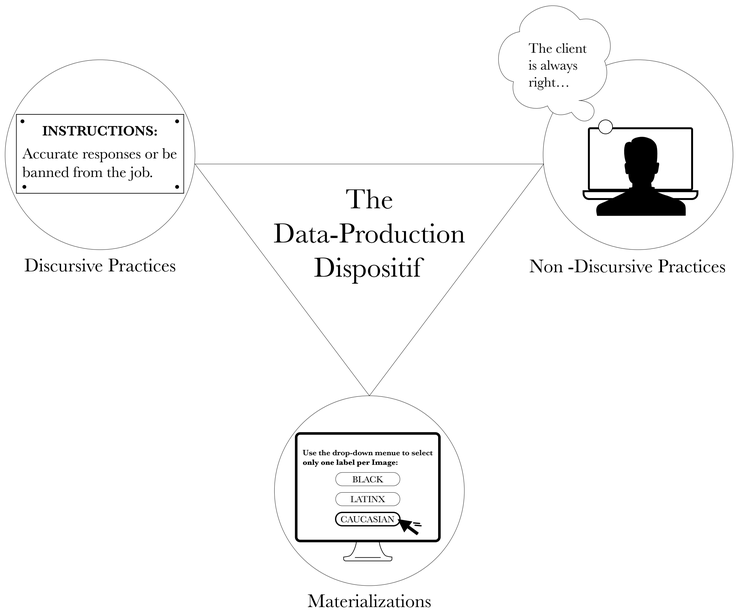
University of Toronto / Weizenbaum Institute

The Data-Production Dispositif

The Foucauldian notion of **dispositif** describes an **heterogeneous ensemble** of elements that shape each other and produce **knowledge** and **power**.

The **data-production dispositif** is thus the network of discourses, work practices, subjects, and objects that enable the (re-)production and circulation of specific discourses in and through data work.

The data-production dispositif responds to the growing demand for data and labor, and **has crucial effects on the outputs that ML models will consider to be true.**



This Research

We explore the experiences of **data workers** of three crowdsourcing platforms in Venezuela and a business process outsourcing (BPO) company in Argentina. Through **dispositif analysis**, we studied the (1) **linguistically-performed elements** (what is said/written in task instruction documents), (2) **non-linguistically performed practices** (how is data work for ML performed and what social contexts inform work practices), and (3) **materializations** (how linguistically and non-linguistically performed practices translate into objects that enable or constrain work)

Findings

Linguistically-Performed Elements

Instructions carry meanings that are **self-evident to requesters** but not necessarily relevant to Latin American data workers. E.g., labeling according to racial categories based on **US-centric conventions**. Taxonomies comprise classifications that prioritize **commercial application** or are easier to **operationalize in computational terms.** Workers have **little room** to improve labels or voice ethical concerns. Instruction documents include warnings that reinforce hierarchical structures and **compel workers to follow orders**

EXAMPLES:

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 thought out and accurate responses**. ***Otherwise, you will, unfortunately be banned from the job :(***

Findings (Cont'd)

Non- Linguistically-Performed Practices

Workers label data according to the **pre-defined truth values** contained in task instructions. They are subject to **control and surveillance** and have **no information** about the ML applications that will be trained on the basis of the data they produce.

Precarized labor conditions as well as the context of **poverty and lack of opportunities** in the regions where data production is outsourced are fundamental for the correct *functioning* of the data-production dispositif. They make workers **dependent on requesters** and, as such, **obedient to instructions.**

Dispositif's Materializations

The hegemony of the pre-defined truth values instructed by requesters is stabilized through **narrow task instructions and other documents** such us non-disclosure agreements. Specially tailored **work interfaces** (see below), and **tools to surveil workers and quantify their labor** serve the same purpose. Managers in BPOs and algorithms in crowdsourcing platforms oversee workers' outputs and make sure tasks are done according to clients' expectations.

These artifacts constitute **dispositif's materializations** as they encode discourses, ensuring their reproduction, circulation, and normalization.