Re-Introducing the Human in AI Labor Terminology with Data Workers' Vernacular Terms

Annabel Rothschild, Carl DiSalvo, and Betsy DiSalvo
arothschild@gatech.edu, cdisalvo@gatech.edu, bdisalvo@cc.gatech.edu

Summary

When confronted with the language AI professionals use to describe the labor that makes AI systems possible, many data workers find that language to be both vague and de-humanizing, such that it obfuscates the myriad and extensive work that must take place. For example, "preprocessing" is used to describe all data work that must take place before a dataset can be used to train a model. But the actual labor involved could take weeks or even months, and range from building scripts to scape a website for data to the tedious work of data standardization. In this position paper, we report on data workers' impressions of key terms in the Datasheets for Datasets project, such as "preprocessing," which, as shorthand, minimize the work entailed. "Preprocessing," here, obfuscates the significance of basic data work. These reflections are part of a larger project in which we are trying to position data workers as early auditors of AI systems, as they are on the most intimate terms with a dataset that will be used to train an AI system. The reflections described here resulted from data workers trying to use the Datasheets for Datasets to document both their labor and how they understood the dataset they were working with, and the troubles they encountered in doing so, leading to the group discussion from which we pull our Findings.

We hope to bring data workers' vernacular terminology regarding data work for AI systems to the workshop, to offer it as a way to get a more precise vision of the human labor required in the early stages of the AI pipeline. From other participants, we are excited to learn about (dis)similar language and how it represents the labor that takes place at other parts of the AI pipeline.

Introduction

What does the field of AI ethics bring to foregrounding the human labor in the AI pipeline? AI ethics brings humans to the forefront by prioritizing the implications of AI systems on human life and experience. However, the way AI ethics gets done in practice is often overly restrictive; who is accorded the "expertise" and "respect" to perform things like algorithmic audits often leaves out the early-stage pipeline workers who perform the labor that makes AI systems possible – e.g., data annotators. This is a problem because, we propose, it is the workers who perform data labeling and annotation that are often best positioned to make sense of what is in the datasets used to train and develop AI systems.

Over the past year, we have been working on a project to expand on the successful "Datasheets for Datasets" (DfD) project [2] to make it a part of everyday data work practice. We have been doing so by incorporating it into a data wrangling tool we've built for Google Sheets, nicknamed *Datum Fieldnotes*. The data work tool is aimed at civic and non-profit data workers and, along with automating cell-change level documentation, also introduces users to the DfD project. However, in our user testing, we noticed something: even though several participants were familiar with the concepts that were covered in the DfD project, the language in which the document is written threw them. Namely, the AI-speak (or terms computational professionals use to describe AI and ML systems) was not only foreign, but participants pointed out that it disguised human effort involved in the various stages.

Our work is not a critique of the DfD project – rather, it follows a budding genre of work that extends the project for specific domains and practices. DfD has proven useful as a tool for critical thinking about the data used to train AI systems [2]. For example, datasheets have been adapted for Speech Language Technologies [1] and for NLP [4]. Our investigation is centered on the vernacular language that data workers use to describe the datasets they work on and the data techniques they perform. What we observe is that the terminology that emerges is much more tied not only to data as a tangible, materially-grounded concept, but it also re-centers the humans (both

the data practitioner and data subject). The vague, disembodied language that comprises AI-speak contrasts sharply with its humanized vernacular cousin. Consider the use of "*instance*," which is used throughout the DfD project to indicate an indeterminate datum (or single constituent of a dataset). None of our participants in user testing recognized "instance" and several asked for a definition of the term in context. All our participants are experienced data professionals; their lack of familiarity is not one of disciplinary shallowness. When we clarified the term in context, participants were quick to elaborate that they understood the term ("ooh that makes sense") but they used different language. Some of the terms thrown around included "rows" or "lines" (for participants who worked mostly with spreadsheet data) but participants also frequently spoke in terms of their data subjects – an instance was "a person," "a place or address," or "a story" (for a participant who deals mostly with oral history transcripts).

These language swaps might seem inconsequential – a person is, of course, an instance of a dataset of individuals – but their discursive implications are significant; the conversations that emerged in conversations with our participants, for example, in terms of data privacy, is tightly coupled with the idea of humanness and the right of an individual to privacy. "Instance" did not provoke the same reaction, since the term abdicates concrete form, and therefore expectations of privacy. Besides 1-to-1 language swaps, participants also saw the need for expanded language. For example, "*preprocessing/cleaning/labeling*" fails to capture all that is required to make a dataset consumable by a hungry AI model. Participants pointed out that depending on the dataset in its original form, something like "*discretization or bucketing*" might not have occurred, but very likely labor-intensive processes of things like "*standardization*" or "*organization*" or "*reconciliation*" did. Data never emerges from thin air into pretty CSV files, our participants countered, rather there is always vast and currently often-unspoken labor that produces those CSVs.

In this position paper, we explore the discordance between data workers' vernacular language and the AI-speak used by computational professionals. We observed this discordance in the process of user testing for *Datum Fieldnotes*, with 12 participants. We then designed and ran a focus group specifically for the purpose of following up on this language discordance. Our research, thus, is still in the early stages. The focus group participants were 5 data workers, who we asked about the language they use for common concepts, with an initial eye towards creating an equivalent vernacular version of the DfD questions (see Fig. 1). As we discuss in the Findings section of this report, however, a number of design questions emerged from this focus group – related not just to language – that challenge the disembodied nature of datasets and re-ground those artifacts in their human creators and contributors (and sometimes subjects).

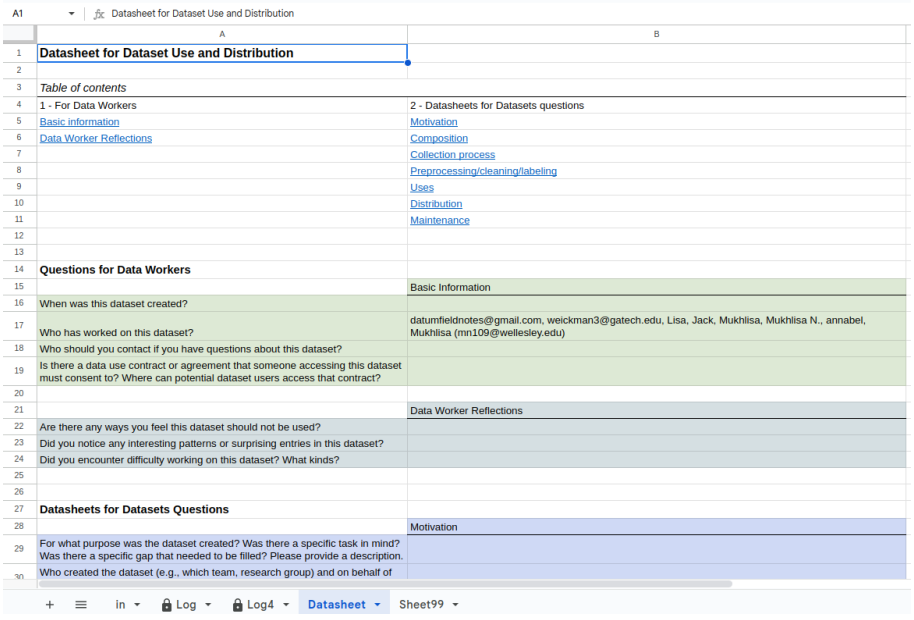

Figure 1: Screenshot of default Datasheets page as generated by Datum Fieldnotes tool within the Google Sheets environment.

In bringing these insights to "The Work of AI" workshop, we hope to explore how language shifts (moving towards applied, vernacular data work terminology) helps us recognize the humans in the loop, especially in terms of the labors that make possible the early stages of the AI pipeline. In other words, we explore the differences between the way on-the-ground data workers talk about their work, vs the way AI engineers do, as a way to chart the topography of early stages of the AI pipeline.

Project Site and Methodology

This work occurred in response to running user testing of a tool, Datum Fieldnotes, one functionality of which is auto populating a new tab in a spreadsheet with DfD questions (see Fig. 1). User testing was conducted with 13 participants: 7 Fellows from DataWorks[1] and 5 civic or non-profit data professionals. Throughout these sessions, when it came to the DfD activity in the session, we asked participants to answer some of the DfD questions with regards to a dataset they'd been using to play around with the tool – all participants were thrown by several terms used, including "instance" (of a dataset), "sensitive data" (sensitive for whom?), "noise" and "redundancies" (noisy data or duplicate entries).

Reflecting on these terms that provoked confusion, we held an hour-long, in-person focus group with 5 DataWorks Fellows (3 of whom had taken part in the original tool testing, 2 of who had not). We asked them a series of open-ended questions, with multiple participants answering each question. We asked three kinds of questions. First, we asked Fellows to describe, in their own words, the kind of work they do, anticipating high level descriptions of the kinds of data work they perform, with the goal of ascertaining how many different kinds of data work are required to "preprocess" data. Second, we asked what word or phrase Fellows would use to describe a particular data scenario (e.g., "How do you refer to a single unit of data?") and, once they'd answered, we asked them how they felt about the equivalent DfD term (here, "instance"). The third category of question was related to higher level concerns about data, for example, about terminology for data that could provoke emotional responses, and whether DfD' use of "anxiety" felt sufficient.

Early Findings and Discussion

While we are still in the process of reviewing and reflecting on these findings, one key dimension – which are excited to potentially share with other workshop attendees – is the way that vernacular language can help point out labor that is hidden across the AI pipeline. Where the parlance of AI designers and engineers is rooted in computational practice and makes use of theoretical terms borrowed from mathematics and physics, to those who aren't familiar with this verbiage, it functions as a kind of *Newspeak* through which verbosity obscures facts. As Slava Gerovich has shown [3], this style of language, especially used in combination with computational systems, has the ability to reinforce social boundaries, such as those who are subject to AI systems and those who design and engineer them, obfuscating the roles of everyone in between, or those with a foot in both camps. Specifically, such language can be used to make artificial notions of objectivity and truth seem reasonable, which is fundamental to contemporary AI practice.

In breaking down the work of AI, particularly in those domains, understanding what language is used, and by whom, can point out the hidden labors (both social, e.g., relationship management, and material, e.g., the range of data work). Further, studying the language of AI developers can help us understand how the AI developer community has come to understand those

[1] DataWorks is a combined work training program, data services provider, and engaged-research vehicle hosted in Georgia Tech's College of Computing. As a work-training program DataWorks provides paid one-year fellowships to adults ('Fellows') interested in transitioning to the tech sector, focusing on all areas of data work (cleaning, standardization, analysis, client communications, etc.) through both dedicated educational modules and on-the-job training via client projects, primarily for civic and non-profit organizations. https://dataworkforce.gatech.edu/

labors. Specifically, the distance that developers may be putting between themselves and the vast array of professionals who make them possible.

References

1. Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6: 587–604. https://doi.org/10.1162/tacl_a_00041
2. Karen L. Boyd. 2021. Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2: 1–27. https://doi.org/10.1145/3479582
3. Slava Gerovitch. 2002. *From newspeak to cyberspeak: a history of Soviet cybernetics*. MIT Press, Cambridge, Mass.
4. Orestis Papakyriakopoulos, Anna Seo Gyeong Choi, William Thong, Dora Zhao, Jerone Andrews, Rebecca Bourke, Alice Xiang, and Allison Koenecke. 2023. Augmented Datasheets for Speech Datasets and Ethical Decision-Making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (FAccT ’23), 881–904. https://doi.org/10.1145/3593013.3594049