Developing Pro-Social AI Training Datasets Through Data Workers' Critical Perspectives

Annabel Rothschild
arothschild@gatech.edu
Georgia Institute of Technology
Atlanta, GA, USA

Abstract

The massive datasets used to train AI models frequently contain offensive and harmful entries that are only caught during the system's later performance. The data workers who curate such datasets are experts in datasets' contents, but silenced by horrible labor conditions and lack of respect by their employers. In my dissertation, I study how to build 1) workplaces, 2) workflows, and 3) tools to elicit and embrace data workers' observations in dataset development. My goals are creating datasets safe to use to train AI systems and developing a more pro-social data labor paradigm.

CCS Concepts

• **Human-centered computing → Ethnographic studies**.

Keywords

Computer supported cooperative work, data annotation, data workers, AI ethics, civic and non-profit organizations

ACM Reference Format:
Annabel Rothschild. 2025. Developing Pro-Social AI Training Datasets Through Data Workers' Critical Perspectives. In *The 2025 ACM International Conference on Supporting Group Work (GROUP Companion '25), January 12–15, 2025, Hilton Head, SC, USA.*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3688828.3699652

1 Introduction & Background

AI and ML systems are increasingly ubiquitous, with recent advances in LLMs and image generators, such as OpenAI's ChatGPT and DALL·E, creating new urgency in future of work conversations [1, 6, 8, 11, 15]. My work explores how the massive datasets used to train these systems, collected and curated by a global workforce of data workers, come into being. Specifically, I examine what the **perspective** and **lived experience** of a data worker contributes to the data labors they perform.

The perspectives of data workers who build the datasets for data-intensive systems, such as AI and ML systems, frequently goes unappreciated. Data workers have a unique on-the-ground view of the dataset and how it has been designed and developed, given that they are the executors of this work. Many of the problems we see with "biased" AI and ML systems can be traced back to issues with the dataset on which the system was trained. Consider the case of *ImageNet*, one of the most impactful computer vision (CV) benchmarking datasets to have been developed, facilitated by the labor of Amazon Mechanical Turk (AMT) workers (Turkers) [3]. The labels Turkers were offered to label images were based on *WordNet* [10], which has been in wide circulation since 2011. These labels, as demonstrated by Prabhu & Birhane, included terms that are offensive and not safe for work (NSFW), along with a host of nonconsensual pornographic terms [2]. Did the Turkers who annotated *ImageNet*'s entries come across these terms? Could they have alerted the *ImageNet* designers to problems with the use of *WordNet* labels before *ImageNet* became a critical benchmark dataset for CV systems?

Having seen the role that data workers equipped with CDL can play in positively shaping datasets, both in technical detail and sociocultural premise, I believe that building healthier, most pro-social AI and ML systems begins with intellectual partnership with data workers in dataset creation and development. My work is motivated by the role that data worker perspective can play when data workers are empowered to practice critical data literacy (CDL), as I observed during my ethnographic fieldwork with DataWorks, a combined work-training program, data services provider, and research platform [4]. CDL goes a step beyond regular data literacy, which refers to a skillset for reading and understanding data statistics and data visualizations [9]. In addition to those skills, practicing CDL requires developing a *critical consciousness* [5], in the tradition of Paulo Freire [16], which means being able to question how these data summaries were arrived at, what might be behind the motivation for their creation, and whom they benefit. Finally, to practice CDL also requires a workplace that supports this critical practice, namely in the form of encouraging workers to speak up and out about problems or concerns they have with dataset development.

My overarching research question (RQ) is: *what is the role of perspective in data work, and how can we incorporate the perspective of data workers as partners in dataset contextualization?* My consequential work answers three subquestions:

- **RQ1**: why do we need better contextualization practices in data work, and what is the current state of data work annotation practices?
- **RQ2**: what is the relationship between critical data literacy and properly localized AI and ML systems?
- **RQ3**: how we can collect and integrate more varied perspectives to relocate our AI and ML systems?

Anticipated contributions: My work facilitates the development of healthier, more pro-social AI and ML systems. My completed and in-progress works are situated in critical data studies,

GROUP Companion '25, January 12–15, 2025, Hilton Head, SC, USA.

ACM ISBN 979-8-4007-1187-9/25/1
https://doi.org/10.1145/3688828.3699652

with an emphasis on building out approaches to the integration of worker perspective in datasets in large-scale dataset development sites.

2 Completed Work

My completed work falls into two categories: building workplaces and workflows to support the integration of data workers' CDL and perspective into the datasets they work with.

2.1 Study 1: Why we need CDL-supportive workplaces

I have explored the experiences of civic and nonprofit workers who perform data work as a large part of their job, but do not identify as data scientists or data analysts [13]. For example, an affordable housing policy expert who works at a nonprofit and spends their days analyzing housing open datasets. Faced with limited time and often little budget, our interlocutors developed extensive processes of collaboration, often with individuals in other departments or even organizations. Further, they found creative and meaningful ways to get "close to the source" of data collection (namely, those who collected the data) when they saw data points that were surprising or potentially low quality. Performing this labor of collaboration and contextualization is critical for responsible AI, and it is work done best not by data scientists or AI architects, but by data workers and domain experts.

2.2 Study 2: Building CDL-rich workplaces to support data work as a profession

We have designed DataWorks to be such a democratic workplace, in which the Fellows have say over what projects they will work on and are encouraged to bring their lived experience and perspective as tools of the trade [7]. In DiSalvo et al. [4], we trace the effects of a CDL course I designed and taught at DataWorks on an enusing client project. The Fellows used their lived experience and critical perspective to change the way the dataset was constructed, fundamentally improving the the resulting dataset in the client's own perspective.

2.3 Study 3: Reporting on the status-quo of data annotation tasks beyond DataWorks

Designing workflows—or actual request for and submission of data work tasks—is considered an open problem, given the frequent (per-platform specific) lack of clarity around quality of work performed [14]. This lack of transparency creates an unequal power balance between workers and requesters. Workers are thus not inclined to trust requesters, resulting in task submissions that are low quality or even fraudulent. This further degrades the working relationship.

Upsetting this power imbalance and building trust relationships between workers and requesters requires increasing transparency on the part of requesters. I have studied how requesters of datasets to be used for training AI systems understand the workers who create and curate those datasets [14]. Among the concerning practices we discovered, requesters use proxies, or empirical tests they believe to ascertain the labor quality of potential workers, as well as verify the task submissions of workers. For example, requesters trace the IP address of workers to verify that they are located in a particular place, even though they know workers can spoof IP addresses with VPNs. Proxies are essentially workarounds for establishing genuine trust relationships between workers and requesters.

2.4 Study 4: Creating transparent & pro-social data work tasks on crowdworking platforms

As part of the critical data literacy course I taught at DataWorks, several of the Fellows tried working on major data work platforms, including AMT. Unlike other studies of how these platforms could be improved, our study is co-authored and co-constructed by workers who have seen alternatives, namely, DataWorks [12]. Combined with the DataWorks operating structure, the experience of being invited to share impressions and opinions, as well as the open-ended nature of the engagement (discussion took place over multiple sessions), the Fellows were not constrained in time nor imagination in considering what a transparent, pro-social data work request looks like. We found that common practices of requesters obscuring their identity – e.g., through use of pseudonyms for requester profile names – while simultaneously demanding deeply personal information from their workers was a prominent cause of distrust. To enhance trust, it appears requesters will need to earn worker disclosure by sharing their own identity and affiliations, which goes against current requester practices.

3 Work in Progress — Study 5: Tools to support data documentation and inquiry in-situ

Inspired by both observations at DataWorks, and by our findings of civic and non-profit data workers' impressive documentation and reconciliation practices, I am developing *Datum Fieldnotes*. This Google Sheets add-on supports these data contextualization practices for spreadsheet-based work. Datum Fieldnotes thus facilitates documentation of data work in settings, like civic and non-profit ones, with fluctuating resources and personnel, as well as supporting data contextualization more broadly through notes integrated with the dataset itself. The tool is designed to automatically track changes at the cell level in an easily-queryable tab (the "log"), within the existing spreadsheet. Further, the tool allows users to mark up either individual changes or values with notes, which, unlike Google Sheet's inbuilt comment functionality, can be highlighted with custom colors (corresponding to types of concerns) and are copied to the log. The log thus also functions as metadata for the dataset itself, and a series of educational materials we've developed show how to use the log to better understand the dataset itself. These functionalities allow data workers to document any concerns or questions they may have about a dataset, as well demonstrate a record of their work for the purpose of professional career experience.

4 Goals for Doctoral Consortium

Having started my PhD during the pandemic, I have not much chance to interact with either peer or senior scholars outside of my institution. I look forward to getting fresh perspective on my work, as well as critical feedback on the data tool I am developing, since the bulk of my work thus far has been ethnography-based, as opposed to artifact creation. I hope, in particular, to get feedback from other GROUP scholars concerned with practice-oriented

work and in-situ workplace studies, particularly for often overlooked groups, such as civic and non-profit organizations. I hope that I can add my perspective as both of a developer and critic of data annotation & data-intensive systems in the workplace to the GROUP Doctoral Consortium, as a site of knowledge sharing and collaborative critique.

Acknowledgments

Thank you to my advisors, Dr. Betsy DiSalvo and Dr. Carl DiSalvo for their support. The work described here is supported by National Science Foundation Award #1951818 DataWorks: Building Smart Community Capacity, a Google collaboration gift — "Examining the data practices of human-in-the-loop ML development", and a Kapor Foundation Dissertation Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any supporters of this work.

References

[1] Ajay Agrawal, Joshua Gans, and Avi Goldfarb. 2022. ChatGPT and How AI Disrupts Industries. *Harvard Business Review* (Dec. 2022). https://hbr.org/2022/12/chatgpt-and-how-ai-disrupts-industries Section: AI and machine learning.

[2] Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision?. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1536–1546. https://doi.org/10.1109/WACV48630.2021.00158 ISSN: 2642-9381.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848 ISSN: 1063-6919.

[4] Carl DiSalvo, Annabel Rothschild, Lara L. Schenck, Ben Rydal Shapiro, and Betsy DiSalvo. 2023. When Workers Want to Say No: A View into Critical Consciousness and Workplace Democracy in Data Work. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (Jan. 2023), 23.

[5] Paulo Freire. 2000. *Pedagogy of the oppressed* (30th anniversary ed ed.). Continuum, New York.

[6] Sam Gilbert. 2023. What might be the economic impact of AI tools like ChatGPT? https://www.economicsobservatory.com/what-might-be-the-economic-impact-of-ai-tools-like-chatgpt

[7] Lilly C. Irani and M. Six Silberman. 2013. Turkopticon: interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Paris France, 611–620. https://doi.org/10.1145/2470654.2470742

[8] Michael Kan. 2023. OpenAI Surveying People on 'Economic Impact' of ChatGPT. *PCMAG* (Feb. 2023). https://www.pcmag.com/news/openai-surveying-people-on-economic-impact-of-chatgpt

[9] Tibor Koltay. 2015. Data literacy: in search of a name and identity. *Journal of Documentation* 71, 2 (March 2015), 401–415. https://doi.org/10.1108/JD-02-2014-0026

[10] George A. Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (Nov. 1995), 39–41. https://doi.org/10.1145/219717.219748

[11] David Rothman. 2023. ChatGPT is about to revolutionize the economy. We need to decide what that looks like. *MIT Technology Review* (March 2023). https://www.technologyreview.com/2023/03/25/1070275/chatgpt-revolutionize-economy-decide-what-looks-like/

[12] Annabel Rothschild, Justin Booker, Christa Davoll, Jessica Hill, Venise Ivey, Carl DiSalvo, Ben Rydal Shapiro, and Betsy DiSalvo. 2022. Towards fair and pro-social employment of digital pieceworkers for sourcing machine learning training data. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3491101.3516384

[13] Annabel Rothschild, Amanda Meng, Carl DiSalvo, Britney Johnson, Ben Rydal Shapiro, and Betsy DiSalvo. 2022. Interrogating Data Work as a Community of Practice. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–28. https://doi.org/10.1145/3555198

[14] Annabel Rothschild, Ding Wang, Niveditha Jayakumar, Lauren Wilcox, Carl DiSalvo, and Betsy DiSalvo. 2024. The Problems with Proxies: Making Data Work Visible through Requester Practices. In *Proceedings of AIES*. San Jose California USA.

[15] Shamika Sirimanne. 2023. How artificial intelligence chatbots could affect jobs | UNCTAD. *United Nations Conference on Trade and Development* (Jan. 2023). https://unctad.org/news/blog-how-artificial-intelligence-chatbots-could-affect-jobs

[16] Alan Freihof Tygel and Rosana Kirsch. 2016. Contributions of Paulo Freire for a critical data literacy: A popular education approach. *The Journal of Community Informatics* 12, 3 (2016).