Data Cleaning, Discard Studies, and Discretionary Power

Pınar Barlas

Faculty of Information and Media Studies, The University of Western Ontario, London, Ontario, Canada
pbarlas3@uwo.ca

Abstract

Data cleaning is an overlooked yet impactful step in the Artificial Intelligence (AI) development pipeline, leading to negative downstream impacts when performed carelessly. Using Discard Studies as a framework, I propose an ethnographic study of how data practitioners exercise their discretionary power during the data cleaning process, particularly with respect to discarded data. The in-depth knowledge of the data cleaning process gained as a result of this study will allow us to improve guidelines and education on data cleaning for more ethical AI development.

Problem Space

When it comes to investigating or mitigating the negative ethical impacts of Artificial Intelligence (AI), it is impossible to overlook the importance of data in the AI development process. Whether for training or testing, datasets are essential to the development of machine learning models. However, datasets can be a source of harmful model outputs, due to representational imbalances and injustices (Paullada et al. 2021) or even the construction of measurements (Jacobs and Wallach 2021; Pine and Liboiron 2015). Therefore, a still-young field of Critical Data Studies (CDS) has emerged to examine the socio-political power of data and its construction process (Iliadis and Russo 2016; Gitelman 2013), following in the footsteps of the well-established Science & Technology Studies (STS) field (e.g. (Bowker and Star 2008; Star 1999)).

One of the main overlaps between the CDS and STS fields is the view of datasets as infrastructure that supports later (often technical) work. Datasets contain subjectivities and uncertainty that get erased (Miceli and Posada 2022; Muller et al. 2021; Plantin 2019) in order to deem the dataset "clean" and build upon it machine learning models. The perception of the "ground truth" dataset as an objective representation of the world contributes to its invisibility.

As researchers have identified dataset construction as a potential site of problems, various studies have looked at dataset documentation—as both a source of information about the construction process (Paullada et al. 2021) and a potential mitigation measure for future harms (Gebru et al. 2021)—but have found the existing levels of documentation superficial and insufficient in almost all cases. By leaving out the data-related decisions from the documentation, the data processing stage is made invisible, and what cannot be seen cannot be challenged. This risks further entrenching the worldviews of the dataset creators in the dataset (Miceli and Posada 2022), which are a potential source of harm.

Datasets are still visible while being built; however, data practitioners often report that data cleaning (processing) is both the task that takes up most of their time, and the least enjoyable (e.g. (CrowdFlower 2016)). Likely because it is a lengthy, 'boring' task, data cleaning is undervalued and overlooked, leading to negative, downstream effects in AI development (Sambasivan et al. 2021). One potential risk is (inadvertently) leaving out or aggregating smaller subpopulations in the data for convenience or "data quality," leading to an exclusion of marginalized people (Simson, Fabris, and Kern 2024).

One way to examine this discarding of data (and the people they represent) is through the emerging subfield of Discard Studies (DS). Researchers looking at waste systems in environmental sciences started studying the wider contexts of what is considered waste, how it is disposed of, and what practices this setup normalizes. Applicable to many areas, DS looks at how discarding practices establish and maintain systems of power, whether the discards are material (e.g. trash), practices, regions, or people (or their data) (Liboiron and Lepawsky 2022). Therefore, the DS lens is fitting to examine data cleaning practices—'getting rid of' data deemed 'dirty' whether by elimination or transformation.

To identify the hidden mechanisms at play, though, it is not enough to examine educational materials designed to teach the fundamentals of data cleaning. Participants in studies on data practices often mention how their education and training is not applicable to real world data (e.g. (Sambasivan et al. 2021)), and that data cleaning practices vary greatly between people despite attempts at standardization (Plantin 2019).

Due to a lack of standards that can apply to all kinds of datasets, and the ubiquity of uncertainty and subjectivity inherent in real world data, it is ultimately up to the individual data practitioner to determine exactly how a dataset should be cleaned. This means each practitioner holds a lot of discretionary power—e.g. the power to determine whether a

rule applies or not in a borderline case (Hong 2023), or how to deal with outliers and 'minority' data. Therefore, data cleaning is undeniably a site worthy of investigation for power dynamics (Sambasivan et al. 2021).

Methodology Proposed

Since dataset documentation is often inconsistent and underspecified when it comes to data practices, and people have unique approaches to data cleaning, the best way to investigate the actions of data practitioners is to conduct an ethnographic study on a team cleaning data for an AI (machine learning) task. By observing the everyday tasks and discussions related to data cleaning, I will be able to identify habits and other taken-for-granted beliefs that participants may not be able to identify when asked directly (Star 1999). Often, data practitioners rely on their 'intuition' and tacit knowledge, of which they may not even be aware (Muller et al. 2021). However, interviews will be useful to supplement the findings by elaborating on the thought processes behind the actions I observe.

As a piece of infrastructure-in-making, dataset construction is a great candidate for ethnography, which is established as a means to "[surface] silenced voices, [juggle] disparate meanings, and [understand] the gap between words and deeds" in the seminal work by Star (1999, p.383). I will apply the methodologies that the Discard Studies authors list: defamiliarization, denaturalization, decentering, and depurification (Liboiron and Lepawsky 2022). These methodologies interrupt what is deemed normal or intuitive, identify the conditions that make discarding the optimal solution, look at the powerful center that remains after the discarding, and investigate the underlying ideology of the discarding practices. In addition, enacting datasets ethnographically may prove to have the same benefits as doing so with algorithms and transform any access issues into findings themselves (Seaver 2017).

Lastly, I will be conducting a short observation at a class discussing data cleaning and hold focus groups with the students in said class, to understand whether the practices I observe in the industry team are reflected in (or in conflict with) the current formal education. Students—as newcomers to a community of practice—are great indicators of what is taken for granted in said practice, since they run into interruptions to their experience (Bowker and Star 2008).

Contributions & Significance

The first and most straightforward outcome of my research will be a detailed description of data cleaning practices, which is currently lacking in the field and difficult to research due to the time and access limitations of ethnographic studies in industry settings. This in-depth knowledge of the practices will also allow me to make policy recommendations for improved, practical ethical AI development guidelines. Finally, I will use the findings to develop pedagogical guidelines, so that formal education and other training on data cleaning may be improved. Together, these outcomes of my research will lead to better data processing pipelines in the future, reducing harm to marginalized groups.

References

Bowker, G. C.; and Star, S. L. 2008. *Sorting Things out: Classification and Its Consequences. 1. paperback ed., 8. print. Inside Technology*. Cambridge, Mass: MIT Press.

CrowdFlower. 2016. Data Science Report. https://www.kdnuggets.com/2016/04/crowdflower-2016-data-science-repost.html. Accessed: 2024-12-20.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gitelman, L. 2013. *"Raw Data" Is an Oxymoron*. The MIT Press. ISBN 9780262312325.

Hong, S.-h. 2023. Prediction as extraction of discretion. *Big Data & Society*, 10(1): 20539517231171053.

Iliadis, A.; and Russo, F. 2016. Critical data studies: An introduction. *Big Data & Society*, 3(2): 2053951716674238.

Jacobs, A. Z.; and Wallach, H. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 375–385. New York, NY, USA: ACM.

Liboiron, M.; and Lepawsky, J. 2022. *Discard Studies: Wasting, Systems, and Power*. Cambridge, Massachusetts; London, England: The MIT Press.

Miceli, M.; and Posada, J. 2022. The data-production dispositif. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–37.

Muller, M.; Wolf, C. T.; Andres, J.; Desmond, M.; Joshi, N. N.; Ashktorab, Z.; Sharma, A.; Brimijoin, K.; Pan, Q.; Duesterwald, E.; and Dugan, C. 2021. Designing ground truth and the social life of labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. New York, NY, USA: ACM.

Paullada, A.; Raji, I. D.; Bender, E. M.; Denton, E.; and Hanna, A. 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns (N. Y.)*, 2(11): 100336.

Pine, K. H.; and Liboiron, M. 2015. The politics of measurement and action. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 3147–3156. New York, NY, USA: ACM.

Plantin, J.-C. 2019. Data cleaners for pristine datasets: Visibility and invisibility of data processors in social science. *Science, Technology, & Human Values*, 44(1): 52–73.

Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; and Aroyo, L. M. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15. New York, NY, USA: ACM.

Seaver, N. 2017. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big data & society*, 4(2): 2053951717738104.

Simson, J.; Fabris, A.; and Kern, C. 2024. Lazy data practices harm fairness research. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 642–659. New York, NY, USA: ACM.

Star, S. L. 1999. The ethnography of infrastructure. *American behavioral scientist*, 43(3): 377–391.