CONSIDER DATA CLEANING

Do you do code work for some data-driven or AI system? How much time do you spend cleaning data?

If you answer "too much", shake your head, or feel kind of frustrated or ambivalent about it, I invite you to:

Imagine a perfect world, with respect to how code/data systems are designed, developed, and maintained. Consider: how much time do you spend cleaning data in this perfect world, and how do you feel about it?

This essay offers neither tools nor processes; it proposes an alternative *feeling* to that familiar sense of "too much" data cleaning. I align this sketch with well-established research on institutional and structural factors at play in the development of code/data systems.

1. DATA WORK

Data requires cleaning because it includes imperfect, human annotation. (A complete lack of human oversight is a problematic goal, though; it can lead to uncontrollable and deeply biased feedback loops [1].)

Critiquing the message of the book "*The Second Machine Age*", Dr. Lilly Irani writes that its authors "ignore the labor of cultural data workers, as if algorithms trained, tuned, and augmented themselves, like magic" [2]. The "training data" demanded by artificial intelligence includes "content moderation" or "curation" tasks and asking "raters" to judge output of automated systems.

This data work is systematically, broadly, and "profoundly undervalued in proportion to the knowledge it helps to create" [3, pp. 180, em. added].

Referencing a different book, "*Mindless*," Irani notes that the problem is not with automation itself, but "with the ways automation entrenches command-and-control relationships between managers and workers" [2]. Controlled and controllable micro-tasks fail to engage with "a wide range of accumulated human ability and wisdom," or "generate more subtle kinds of value" [2].

The perfect world includes not only credit for data work, but also a re-imagining of how this work can better inform the attendant code work.

2. BETTER REVISION

Data work matters, but data work produces imperfect data, which leads to the uncomfortable idea that most of us are doing "too much" data cleaning. Part of this is the often-implicit expectation that it is a pre-amble.

Data cleaning is, in practice, not a one-time pre-processing step, but a recurring process, because repeated revision, of both data and code, is inevitable.

All contemporary AI systems cannot exist without data work; the myth that code/data systems can be unbiased is only possible because this necessary data work is intentionally hidden from view. It also contributes to the myth that a code/data system can "just run" without repeated interrogation and revision.

In "*Algorithms of Oppression*" [4], Dr. Safiya Umoja Noble details example after example of search engines being described as neutral and/or inscrutable; then, after sufficient pressure, being reviewed and found to be biased. Revision is not only possible but done; and often. Despite initial assumptions, revision can and does result in improvements.

In a perfect world, recurring revision arises without external pressure: not as an exception, but as the routine. Data work extends beyond data labeling to finding new areas for improvement through revision.

Imagine: the repeated process of data preparation, including both review and cleaning, is a recognized part of necessary code work, supported by tools and processes. Data preparation, with more time and attention, can adapt to more complex (and more human) data collection and feedback mechanisms.

Data cleaning is unavoidable! Each round of repeated care of data and code is an opportunity to invite new perspectives to code/data technical objects.

In a perfect world, I spend a lot of time on data preparation, and I feel great about it! Instead of adversarial unease about imperfect labels, I feel curiosity. Instead of trying to minimize time spent cleaning data, I use preparation tasks to interrogate the data and its use. Bias is not gone from the system or its context; but it is a subject of active, enthusiastic work.

[1] O'Neil, Cathy. Weapons of Math Destruction: How big data increases inequality and threatens democracy. Broadway Books, 2016.
[2] Irani, Lilly. "Justice for 'Data Janitors'." *Public Culture 15* (2015).
[3] D'Ignazio, Catherine, and Lauren F. Klein. Data Feminism. MIT Press, 2020.
[4] Noble, Safiya Umoja. Algorithms of Oppression: How search engines reinforce racism. NYU Press, 2018.