SAFE FROM THE START:
DEVELOPING PRO-SOCIAL AI TRAINING DATASETS THROUGH DATA WORKERS' CRITICAL PERSPECTIVES

A Dissertation
Presented to
The Academic Faculty

By

Annabel Rothschild

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Georgia Institute of Technology

Georgia Institute of Technology

May 2025

SAFE FROM THE START:
DEVELOPING PRO-SOCIAL AI TRAINING DATASETS THROUGH DATA WORKERS' CRITICAL PERSPECTIVES

Approved by:

Dr. Betsy DiSalvo
Advisor
School of Interactive Computing
Georgia Institute of Technology

Dr. Carl DiSalvo
Advisor
School of Interactive Computing
Georgia Institute of Technology

Dr. Shaowen Bardzell
School of Interactive Computing
Georgia Institute of Technology

Dr. Ellen Zegura
School of Computer Science
Georgia Institute of Technology

Dr. Richmond Wong
School of Literature, Media, and Communication
Georgia Institute of Technology

Dr. Lauren Klein
Departments of Quantitative Theory & Methods and English
Emory University

Dr. Ding Wang
Google Research

Date Approved: April 2, 2025

To the newly-forbidden words (Diversity, Equity, Inclusion, DEI, Inclusive Excellence, Justice): may your memory be a blessing. You're what makes an academic community a fascinating and fun place to be.

We'll meet again!

Atlanta, GA, March 2025.

ACKNOWLEDGMENTS

I am very grateful to my advisor, Betsy, for striking nicely the difficult balance between letting me develop as an independent researcher and shepherding this dissertation down a productive path. Thanks also for all the advice and hot takes (even if almost never related to what we were supposed to be discussing). Her husband[1] also contributed intellectual lineage to this project, when not busy distracting his zealously-focused grad students with random thoughts in #general. I'm not sure I would term Scout DiSalvo a muse exactly, but her contributions (if indirect) to this dissertation are certainly to its benefit.

I am particularly appreciative to all of the DataWorks Data Wranglers & Fellows past, present, and future for their generous participation and patience over the last five years. Special thanks in particular to Dana and Justin, who joined DataWorks shortly after me, and now not only run the show, but do so with enviably good humor and cool heads. It's been a lot of fun to learn with and from you.

Great thanks to my committee members who have provided a wealth of advice in various forms since this dissertation started to take shape: Drs. Shaowen Bardzell, Ellen Zegura, Richmond Wong, Ding Wang, and Lauren Klein. Ding has been both a frequent collaborator, and fantastic mentor who launched me into the world of (critical) data work studies. Richmond and Lauren helped me push my thinking over the years in the social and cultural histories of computing and data cultures; chapters 3 and 5 are, in particular, stronger for their wealth of expertise. Shaowen and Ellen have provided not only advice on the dissertation, but also shared thoughtful conversation on careers in HCI and computing.

A number of other people have generously supported my research. My undergraduate research assistants—Will, Grace, and Mukhlisa—for both putting up with me learning how to be a supervisor and making significant contributions to the work described here. Drs. Eni Mustafaraj, Torben Wallbaum, Ada Lerner, and Janne Lindqvist, and the wider Wellesley

[1] Also known as "Carl," he is officially the co-advisor of this dissertation.

CS department have provided key opportunities for growth and continued support. Georgia Tech's ILL Staff managed to obtain all of the arcane books I requested. Labmates in the CAT & DEAR labs have supported my work in various ways: Michael, Lara, Idel, Kayla, Sarah, Anh-Ton, Britney, Sara, Alyssa, Catherine, Ishita, and Ashley. Peace out, party people. I am grateful also to my many co-authors for their collaboration, which is denoted in chapters to which they contributed.

Much appreciation to friends at GT, including Rachel (*viva la Haus!*), Maddy, Amy, Adrian, Grace (Guo and Barkhuff), Xander, Bea, Jay, Inha, and Soonho, and beyond, including SDS & Christie, Erik J. and his mother, the Wallbaums, Sarah, Dominique, Adam, and Rongjun. Particular thanks to Anh-Ton and Grace for the goss, to Catherine for kaffee und kvetching, and Ashley for good paper recommendations and infectious optimism.

To my parents, thanks for the occasional burritos—turns out there are some perks for staying local for graduate school!—and grand-plant-ering. Thanks also for all the research anecdotes, e.g., *"So I open the coat closet, and under a pile of hats and scarves, there's an unopened Google Nest and my mom swings around the corner, 'You're going to be* so *proud of me!"'* My gratitude to the Coffee Crowd and Uncle Brad for boundless support, to Ben for hiking escapes with Sleuthy. Special thanks, Los, for care during three surgeries and two substantial illness in these five years. And with much love to both of my grandmothers, for cultivating an appreciation of reading, art, adventure, and independence that has shaped both me and this dissertation.

Mit freundlichen Grüßen und vielen Dank, Tack så mycket, Kiitos paljon, Danken dir zeyer fil, et merci beaucoup,

Annabel

This work was supported by (but does not necessarily reflect the views of) the National Science Foundation award #1951818 DataWorks: Building Smart Community Capacity, a Kapor Foundation Dissertation Fellowship, and a Google Collaboration grant.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

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, 2, 3, 4, 5]. 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) [6]. The labels Turkers were offered to label images were based on *WordNet* [7], 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 [8]. 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, more 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 [9]. CDL goes a step beyond regular data literacy, which refers to a skillset for reading and understanding data statistics and data visualizations [10]. In addition to those skills, practicing CDL requires developing a *critical consciousness* [11], in the tradition of Paulo Freire [12], which means being able to question how these data summaries were arrived at, what might be behind the motivation for their creation, and to whom they offer benefit. Finally, to practice CDL professionally 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?* The work that emerges is thus a study of why do we need better contextualization practices in data work, and what is the current state of data work annotation practices? What, then, is the relationship between critical data literacy and properly localized (or, contextualized) AI and ML systems? And finally, how we can collect and integrate more varied perspectives to relocate our AI and ML systems?

Contributions: My work facilitates the development of safer, more pro-social AI and ML systems. Situated within critical data studies, the work described in this dissertation builds out approaches to the integration of worker perspective in datasets developed at archetypal production sites.

CHAPTER 1
INTRODUCTION

1.1 Motivation

Among the unbridled AI hype of recent years, there has been considerably less attention paid to the people who make AI possible. I am not talking about data scientists or ML developers who are frequently fawned upon [13], but the data workers who collect, clean, standardize, and annotate the datasets used to train and refine AI and ML systems. On the contrary, these mission-critical workers have been maligned as sources of "bias" in data work, or that the data-intensive systems they contribute to are somehow worse off for expressions of personal understanding and knowledge.

In *critical data studies* [14] we know that the opposite is true; namely, that it is the conditions of labor set forth by requesters, as well as the strictures and directives with which they instruct workers, that result in what we understand as bias [15]. In fact, what we call bias is really just *perspective*, which is inherent to any human-in-the-loop system. Perspective cannot be removed, but it can be explained and accounted for. In fact, identifying and embracing perspective actually enriches the dataset produced, by way of providing dataset provenance and contextualization. However, our AI and ML systems, as ones meant to generalize over large swaths of people and situations, reflect only the perspective of a handful of the people involved in the development of these generalizable systems—the model designers and developers.

My dissertation is a proposition for a paradigm of AI system development that is sensitized to perspective. Specifically, the auditing role that data workers should play, in addition to the labeling ones they already do. Data workers are the individuals *closest* to the dataset; in other words, they review the individual constituent elements of the datasets they work

with in ways the model development team usually do not. Given that the development team writes the rules for how the dataset should be worked on (e.g., labeled or annotated), data workers are best poised to point out errors either in those directives or concerns with how they might be applied. We can look to serious errors in common benchmarking datasets, such as *ImageNet* and *WordNet*, as examples of when data workers do the best with what they have, but model developers, as dataset 'requesters' have made mistakes or miscommunicated their intents.

The importance of when these generalizable systems fail cannot be understated. Consider the implications of something like commercial facial recognition technology. Such systems have been demonstrated to perform less accurately for faces with darker skin tones and women, and particularly women with darker skin tones [16]. These are commercial systems that have already been deployed; they have resulted, among other deleterious effects, in innocent people being arrested for crimes they did not commit [17].

I posit this mode of dataset production, where the experiences and perspectives of data workers is ignored or even minimized, represents a paradigm of *unsafe* dataset development. An *unsafe* dataset is one that the designers of AI systems who employ such a dataset to train or refine an AI or ML system, do not fully understand and are unable to contextualize. In other words, the dataset has not been audited; there is no demonstrative proof of what is or is not included in the dataset, ranging from concrete objects to social values. In the case of a dataset that is, for example, scraped from the Web, we can immediately see the problem that unverified content can produce [6, 18]. Employed for the purposes of training or refining a generalizable system, e.g., an AI or ML system, an *unsafe* dataset can easily produce results that we do not anticipate, ranging from those that are merely surprising, to those offensive and harmful, as in the case of *ImageNet* [19, 16].

Requesters of data labor have the opportunity, thus, to embrace data workers as professionals highly capable of acting as dataset auditors. I term this the *data-worker-as-dataset-auditor* role. These dataset audits can reveal issues that will lead to the failure of a

generalizable system, because of their firsthand exposure to, and knowledge of, a dataset's contents, gained through their labor (including any number of collection, cleaning, standardization, organization, and labeling). Unfortunately, this does not appear to be the path most requesters choose. Instead, we hear from designers of these generalizable systems that it is actually the *data workers* who are at fault for the failures of these systems to generalize. This is because data workers are alleged to have inflicted these "biases" into their work. Occasionally, system designers fall back on the idea that it is merely the AI system being trained off of "bad" or "biased" data that results in the failures to respect all human populations. Typically, data workers are blamed in part for these discrepancies, even though, under the current labor paradigm, any concerns they might have raised are unlikely to have been accepted [20, 21, 22].

Given the sheer scale of the datasets used to train these models (at the commercial level, often numbering in the hundreds of thousands or even millions), it is difficult to imagine that the model development would be able to review all constituent elements. However, given that many of these datasets are assembled by data workers on digital pieceworking platforms [23] (hereon, "data work platforms"), we once again encounter a group exceptionally well poised to act as reviewers of the dataset: *the data workers themselves*.

In suggesting that we no longer see bias as avoidable and extractable, I argue that we can partner with data workers, to employ their *perspective* as a means of making sense of what exactly is in the dataset at hand. Unlike perspective, *bias* is seen as an unfortunate additive to a system, something that can be rooted out—e.g., a system should *unbiased* against different human populations. It is true that all systems are biased, however, because of the negative connotations of bias, it is not taken *a priori* in the way that perspective is. When bias is seen as an additive, it becomes a problem to be solved after model development, instead of being taken into account from the start. In fact, Browne et al. found that bias, as a term, "threw" AI workers, creating confusion about what the problem was, meaning that it actually impeded these same workers as they tried to reason about more ethical AI

initiatives and systems [24].

Other efforts to negotiate how to address bias tend to conclude in conversation between stakeholders, where the realization is that more stakeholder perspectives are needed, e.g., Lin & Jackson's study of bias-as-error becoming a site for collaboration and mutual learning [25]. For these reasons, I present my work not as AI or ML system "de-biasing", which is a post-hoc (or, reactive) attempt at fixing either dataset demographics [26] or technical manifestations of existing social systemic biases [27].

Instead, I take a proactive approach in which attempt to integrate diverse perspectives from the beginning of system development, starting with the collection and curation of training datasets. In part, the choice of "more perspectives" as opposed to "de-biasing" is also because it stops AI and ML systems from taking on new and novel characteristics, when the issues that we see in these systems are familiar. Perspective is not radically different than positionality, and indeed, in many cases where I use perspective, positionality could be substituted with little difference made.

However, positionality tends to be a reflection on privilege [28, 29], which in this case, would be proximity to, and understanding of, AI and ML systems. What perspective offers that personality does not is the idea that outsider viewpoints are critical—namely, what we want from data workers is their review of datasets from where they stand, which is less prejudiced toward proximity to power than positionality. Further, fostering (critical) perspective as a product of participatory approaches is a known topic in science and technology (STS) spaces [30]. In calling for not only more perspective, but the embrace of individual perspective itself, I am also following Haraway's notion of situated knowledge, and that any individual only holds partial perspective [31]; if we want to get a more expansive understanding of a dataset, we need to employ numerous perspectives, and from alternate sources that differ from the limited perspectives we already have (namely, system developers).

As an example of the contrasts between bias and perspective, consider the infamous

case of the COMPAS (Corrective Offender Management Profiling for Alternative Sanctions) system, which is documented to treat Black defendants as higher risk than their White peers [32]. We have long known about problems of racial equity and equality in the U.S. law enforcement establishment [33, 34]. The COMPAS algorithm is not uniquely racist; instead, it builds on a longstanding history of racism in U.S. recidivism policy and law enforcement establishment. These histories of injustice are merely forward-propagated through datasets of past cases in which they are embedded.

By calling COMPAS biased, we abdicate responsibility for racist decision-making to a mysterious technical "glitch," rather than acknowledging it is a series of judgements consistent with longstanding systemic inequality (particularly with regards to race). Rather than "de-biasing" the COMPAS algorithm—once the system has been deployed and lives have already been affected [32]—we should have consulted a greater variety of individuals and stakeholders when developing the system. It is highly likely that legal scholars, particularly those familiar with CRT (critical race theory), would have pointed out the problem with basing contemporary recidivism policy with the problematic decisions of decades past.

In choosing to adopt this language of *perspective* instead of *bias*, I hope to avoid giving AI and ML systems this rhetorical novelty. The "bias problem" of AI and ML systems [35] is both a buzzword and a means of disavowing responsibility for something that we should understand as an inherent feature of such systems [36]. Bender et al. describe AI (specifically LLMs) as "stochastic parrots," to reify that they are statistical models without human sentience or thought processes [37]. By calling for more (diverse) human perspectives, I aim to similarly reify that these systems are not humans nor think like them—instead, such systems are technical products that merely reflect the existing societal relationships and predilections of the majority world. In short, my stance is consistent with Suchman's critique of AI's "thingness," or the *a priori* acceptance of AI as a novel, personified technology [38]. We should not allow rhetorical agency to AI, because it is not a sentient agent.

The design of generalizable systems, which abstract over local context such as language, culture, and social dynamics, advance a set of perspectives which are historically limited to model developers and perhaps the organizations that employ them. Perhaps the most important way we can diversify these limited perspectives, however, is integrate the knowledge and lived experience of the data workers who create and process the datasets necessary for these systems. This is where the paradigm of *data-worker-as-dataset-auditor* enters the picture.

My dissertation comprises a series of work that begin with demonstrating how the perspective that data workers bring to their work fundamentally shapes the datasets produced by that engagement. By embracing, rather than attempting to mitigate, these augmented datasets, we can anticipate limitations to the use of the dataset, based on data workers' own experiences with and understandings of a dataset. Said differently, this format of data work, I argue, enables the development of datasets that have undergone a basic level of content auditing, as well as shape and format analysis by the people best placed to do that work, or, data workers. In this dissertation I demonstrate how perspective affects data work, specifically how generalizable systems can benefit from more varied perspectives, with a focus on changes to the dataset itself and, subsequently, the systems that are trained and refined through that dataset. I design programs, tools, and interventions to support data workers in channeling their lived experience and perspective into data work, through the development of critical data literacy [12].

1.2 Research Goals and Audiences for This Work

1.2.1 Research questions

The works reported on in this dissertation represent my overarching research question:

> **RQ: what is the role of perspective in data work, and how can we incorporate the perspective of data workers as partners in dataset contextual-**

ization?

In the following chapters, I examine this question from both the data work ethnography site of DataWorks and how learnings, particularly related to critical data literacy, can be transferred to the platform-mediated work site presented by major players such as Amazon MTurk. More broadly, I trace the changes made to a dataset as the result of perspective and then explore ways that the integration of varied perspective can better shape our data-intensive systems, namely generalizable systems, such as AI and ML.

In the following chapter, I explain why I think the role of perspective in data work can bring us closer to best data use practices. This chapter answers the question of *RQ-A: why do we need better contextualization practices in data work, and what is the current state of data work annotation practices?* In the third chapter, I ask *RQ-B: how did we arrive at this current status-quo and whether we have an alternative path for dataset development?*.

To answer these, I turn first to the entangled and mutual history of data science and AI, stemming from the post-WWII cybernetics legacies. I introduce *cyberfeminism* (see 1.3.1) as an alternative ancestor for these systems. I question how cyberfeminism's emphasis on plurality and complexity would change the way our contemporary AI systems look and behave, compared to the products we encounter today, replete with their cybernetic inheritances. Would we thus encounter a fundamentally different paradigm of AI infrastructure?

Second, I explore alterative dataset development and contextualization practiced outside of commercial data science: namely, that of civic and non-profit (CNP) data workers, focused on data production and analysis that focuses on social benefit or social-services organizations. In the fourth chapter, I turn to the role that perspective plays in determining fair and safe use of a dataset, as inspired by those CNP data workers, namely, *RQ-C: what is the relationship between critical data literacy and properly localized AI and ML systems?* In the fifth chapter, I address *RQ-D: how we can collect and integrate more varied perspectives to re-locate our AI and ML systems?*, honing in on the role of data workers on large-scale digital data work platforms, such as Amazon MTurk. In the fifth chapter, I

explore ways to transfer these learning to platform-mediated work, asking *RQ-E: how can we ameliorate the workplace experience and requester-worker power dynamics that define platform-mediated data work?* with an eye towards the role of requesters.

1.2.2 Audiences

Given my goals in this work, there are three groups of individuals likely to encounter this work. The first is other critical data studies specialists and enthusiasts, who may find utility in the detailed descriptions of data work mechanics, particularly those emanating from the DataWorks environment. The other two groups are both AI practitioners (designers, engineers, and developers) who may be either skeptical or intrigued by the ideas presented in this work. For those in the former category, I hope that they will find the investigation of concrete dataset development in the *data-worker-as-data-auditor* model promising and perhaps even a superior option to their status quo. My desired effect on these readers is to convince them that a better paradigm of dataset development is possible, but requires changing the way we, as AI specialists, engage with those who are earlier agents in the AI supply chain—namely, that we must collaborate with data workers as professionals and experts of in their domain and accord them respect as fair employment as such. For those AI practitioners who may be empathetic to the mission of this work, I hope they will find the sequence of work described in this dissertation a useful blueprint for either modeling similar improved AI supply chains within their own organizations.

In the final section in this chapter, I suggest chapters that may be useful to members of each of these three audiences.

1.3 Theoretical Framings and Commitments

The overarching framing of this work is that of *cyberfeminism* (1.3.1), as a multidisciplinary investigation of humans, the Internet, cyberspace, and technology more generally. Cyberfeminism is a flexible, method-agnostic approach which defies easy categorization

and scoping; the classic definition of cyberfeminism itself is in fact a set of one-hundred anti-theses by the old boys network[1]. Readers of this dissertation may find the assortment of theories, methods, intervention styles, projects, and sites challenging—I appreciate their efforts to navigate through the ideas contained. However, rather than a grab bag sourced from epistemology's junk drawer, this plurality is meant to acclimate the reader to cyberfeminism's embrace of complexity and confusion; no interaction is easily defined, nor interchangeable with another. If our subject of study is that of humans, hardware, organizations, and relationships (and the AI supply chain is all of these things) a cyberfeminist approach requires that we approach each candidate of study with the techniques and tools suited to that candidate and no other.

For the sake of readability, I describe here, along with cyberfeminism, two other theoretical frameworks that shape the direction of this document. In particular, *data feminism* and *cosmolocal* theory offer ways to make sense of the many study subjects included in this work in combination. Data feminism offers a means of social comprehension of the power dynamics of which the AI supply chain is replete. Cosmolocality, on the other hand, is an organizational lens on how the coherence of people, place, and technology works at each site of dataset production.

1.3.1 Cyberfeminism

Cyberfeminism is a generative perspective in which all components of cyberspace, along with the people who use it and the Internet infrastructure that facilitates it, are equally subject to the researcher's focus. Paasonen describes cyberfeminism as a "broad range of practices," the practitioners of which need not be necessarily self-identifying as feminists [39, p. 61]. Instead, invoking the work of Haraway among others, Paasonen articulates cyberfeminism as a foil to cybernetic practice, making the former a critique of the techno-culture and command-and-control preoccupation of the latter.

[1] https://obn.org/obn/reading_room/manifestos/html/anti.html

Building off the legacy of researchers like Sadie Plant, cyberfeminist practice rejects the idea of cyber-products as instruments and tools for the predominantly masculine technoculture, instead embracing human interactions with technology (specifically cyberspace) as embodied [40], sensual, and messy [39, p. 70]. Here, cyberfeminism invokes the non-developer as user and interlocultor, of equal enfranchisement to the developers. While data feminism certainly requires identifying and critically considering the affectee(s) of specifically data analysis and infrastructures, cyberfeminism challenges us to think about interactions pertaining to the Internet (or cyberspace) more broadly, pulling from fields including philosophy, gender studies, computing, and more [41].

Invocations of cyberfeminist theory and praxis have been used extensively to make sense of turning points in the history of humanity and computing, particularly regarding gender, and more recently, race [42]. Additional work has sought to chart the ever-evolving cultural understandings of what feminism means, including adaptions of cyberfeminism congruent with third and fourth wave feminism that are more cognizant of intersectionality [43, 44, 45]. Gillis, for example, proposes a cyberfeminist agenda that is not only a celebration of women's experiences with cyberspace, but the challenges that come with embodied experiences of gender and other body politics [46].

1.3.2 Data feminism

Data feminism, as characterized by D'Ignazio and Klein [47], is a way to understand, reflect upon, and critique the larger practice of data science and data-intensive systems (e.g., AI [48]). Following the idea that data is a sociocultural product, and that basing analysis or system-building off of it represents both a technical and cultural practice, data feminism centers an intersectional approach. Core to the practice is studying the manifestations and implications of power in both data work and the larger practice of data science. Who builds data-intensive systems and performs data-centric analysis shapes the way these products or insights come to be understood, as much as how they are constructed.

Further, data feminism invokes the idea of *invisible labor* [49], which often characterizes the work of collecting, cleaning, and preparing datasets for use in larger systems—or the data work performed by the practitioners that are the focus of this dissertation. Another central point of data feminism is the inability of complex sociocultural phenomena to fit into rigid hierarchies and binary classification; performing this discretization or bucketing is (unsurprisingly, for etymology-inclined) a discretionary act that should be understood as such, rather than a straightforward, simple decision. I invoke the concept of data feminism throughout the following chapter in order to convey the role that data workers' personal perspectives and lived experiences can inform both the way datasets are produced, but also how they should be best used.

I am certainly not the first to employ data feminism as a theoretical lens to make sense of data work as a genre of labor. The care and discretion that shapes data work at its best has been examined in the context, for example, of nonprofit data workers [50] and those in business processing organizations (BPOs) [21, 51].

Later in this work I will utilize the concept of *vernacular* language to surface the local data cultures. Stauff et al. delineate the notion of a *data vernacular* as a "small-scale use of operational data by individuals and organizations," separating the vernacular from more strongly "formalized and institutionalized" applications [52, p. 32]. Data vernaculars can clarify both unique features of a given data work site, but also provide us with contrasts to overarching notions of data as practiced by later-stage (in the supply chain) AI professionals. I use the lens of vernacular language as a tool to conduct a larger, data feminism-informed investigation.

1.3.3 Cosmolocal

The concept of *cosmolocal* is inspired by Tsing's critique of capitalism. In her work on scalability and nonscalability, she develops the idea of *nonsoels*, which are non-social landscape elements that can be treated as interchangeable units [53]. These nonsoels are

essential to scalability. The cosmolocal, as a place-based element that cannot be treated as interchangeable, is antithetical to nonsoels.

Cosmolocal thinking has most often been applied to issues of sustainability and production, where the globalization of knowledge is paired with distributed making [54]. There are three core propositions of production under the cosmolocal model: production should be localized to sites of need, generated knowledge should be shared between all participating sites of distribution, and resultant property and governance should be ruled by the 'generative' model whereby all producer and stakeholder communities have a voice [54, p.37-38].

While nonsoels are interchangeable, the sites of cosmolocal production are anything but. To transfer knowledge or collaborative practices between one site and another, there is inherent friction in exchange between different sites of cosmolocal production. Proponents of cosmolocalism posit that it leads to sustainable production and organizations; those that remain flexible are less likely to to be discarded in the long run, as demonstrated by biological systems of adaptation [55]. Cosmolocalism's roots are in sustainable development practices, designed to reflect and mimic the natural adaptations of living systems.

In this dissertation I draw on the cosmolocal in the later chapters especially, to explore why transferring what we have learned from DataWorks should not be subject to a one-to-one transfer on other sites of dataset production. Each data labor site has its own paradigm and power relationships. While the conceptual findings from DataWorks (such as democratic workplace and critical data literacy to support conducting data labor) can and should be applied, they should be formatted in a way that is appropriate to that specific site.

Determining what is appropriate should be done in consultation with the data workers at each site and each determination template may not be applicable to any other site. While the generally participatory design and engaged workplace models worked for development of these values at DataWorks, different instantiations of co-development may be needed when moving beyond DataWorks. For example, on Amazon MTurk, the psuedo-anonymous

model of engagement between requesters and workers has its pros and cons and these should be taken into account when trying to reformulate a model of trust between the two. Inherently, that will look very different than the local manifestation at DataWorks.

1.3.4 Research commitments

No researcher can set aside the political, philosophical, and epistemological commitments that they hold and which shape their work. Throughout the rest of this document, I am preoccupied with making sense of individual perspective in conjunction with contextualizing a dataset; it is only fair that I attempt to characterize my own perspective with regards to the pages that follow. This work is grounded in partnership with DataWorks (the organization and its staff) and after five years of observations and engagement, I adhere with some degree of absolutism to the idea that data workers, as both individuals and professionals, hold valuable knowledge that is necessary to genuine responsible AI (R-AI) practice. Where contemporary R-AI practice often favors those who appear much later in the AI supply chain—namely those closer to model development, often at either high-tech companies or policy bodies—I am much preoccupied with (and frankly more optimistic) about the potential contributions to the field by those currently marginalized in the AI supply chain (or, data workers).

1.4 Contributions

This first portion of my work occurs in the context of DataWorks, as a unique data services provider and engaged research vehicle, where testing and studying alternatives to the paradigmatic status quo for platform-based data work is possible. However, to transfer these lessons and implications for design to platform-mediated data work (e.g., that on Amazon MTurk and competitors), is not a simple one-to-one mapping; the second portion of my work in this dissertation explores the power relationships that must be re-balanced in order to create a workplace environment that is conducive to the pro-social understanding

of data annotators' perspectives and experiences [51, 56].

To this end, I contribute the design and deployment of a data documentation tool, Datum Fieldnotes, to show how these annotators can be formalized and archived, and become actionable *paradata* [57]. The utility of paradata, or metadata about the development process of a dataset, to responsible AI, and development of data-intensive systems more broadly should not be underestimated; Trace & Hodges identify paradata as a view into the assemblage of an entire algorithmic apparatus [58].

My work includes examinations of the power dynamics that shape dataset production, and specifically the previously under-studied role of data work requesters in shaping those engagements, along with the provocation of automatic per-datum (or per spreadsheet cell) transformation documentation. I subsequently present a workable alternative to the status quo of dataset production for training and refining AI and ML systems. This alternative is one in which we produce datasets that are well understood and contextualized, given the auditing work of data workers—specifically, data annotators—in the incorporation of their lived experience and professional understandings of the nuance and nature of data-intensive systems. Throughout this work, well contextualized datasets also indicates datasets that have not been subject to extensive *context collapse*, in which datasets are used and analyzed in a setting divorced from that in which they originated [59, 60].

Both data feminism and cyberfeminism are congruent with the cosmolocal; all orient us towards practices of noticing differences between sites and experiences and acting accordingly. None of the three suggest that a one-size-fits-all approach is appropriate for any given context.

1.5 Dissertation Overview

The rest of this proposal is structured as follows. In the next chapter (Chapter 2), I explain the motivation and background for my research work, stemming from an experience with DataWorks that made realize what the need for CDL is in data work and how we could

develop it at DataWorks. I compare this with the larger context of data work, as it happens more commonly, particularly on quasi-anonymous digital platforms, such as Amazon MTurk. Chapter 3 is an exploration of data work as a cultural practice. The first half of the chapter is a study for how data contextualization happens in practice and the second half traces the marginalization of that work through the universalist impulses of early cybernetics practitioners. In Chapter 4 I turn to how perspective shapes data work, or tracing how the changes in perspective change a dataset itself. Following, Chapter 5 takes a step back and looks at how we can incorporate more (in both quantity and diversity) perspective in data work, turning particularly to data workers as partners and collaborators in this, which requires more pro-social treatment of data workers, looking beyond DataWorks as a unique data services provider. Transferring these learnings it requires rethinking how we engage in requester-worker relationships, with regards to data work tasks. I expand on the challenges of this transfer in Chapter 6, turning to the notion of cosmolocal as one promising approach.

Readers looking for a case study of how contextualization and worker empowerment lead to better datasets should consider the second and fourth chapters. For those interested in sociocultural aspects of data work, the third, fourth, and sixth chapters may offer insight.

CHAPTER 2

BACKGROUND AND RELATED WORK

2.1 DataWorks and Early Beginnings

In 2020, I started both my PhD and my tenure with the DataWorks organization. Functioning as a work-training program, a data services provider, and research platform, DataWorks is a social enterprise run out of Georgia Tech's College of Computing. In its current form, DataWorks hires cohorts of professionals (*Data Fellows*) interested in getting into the field of data work and, data analysis more broadly. In the first cohort, the professionals who joined the organization were known as "data workers" with open ended appointments, while subsequent cohorts were known as "Data Fellows" and their appointment term is a 1-year fellowship.

Under the current program structure, Fellows are hired for a year to hone their skills as data workers through a combination of on-the-job learning on live client projects, as well as explicit learning modules on various aspects and tools of data work. This a change from the first cohort of workers hired by DataWorks; those workers had an open-ended appointment and held the title of *data worker*. Throughout this work, I will alternate between using Fellows and data workers in discussion of the DataWorks project; this is in keeping with when the professionals being described joined the organization.

The Fellows (and the data workers before them) are paid a competitive hourly wage and are fully University employees, meaning they have access to traditional benefits such as health insurance and paid vacation and sick leave. Many of the professionals who join the program are in the process of changing careers, while others are at the very beginning of their careers. None of the data workers have a traditional background in technology or data services, and most hold a GED or high school diploma as their highest academic

credential. The goal of DataWorks as a work-training program is not only to help the Fellows continue to build their skills, but also to serve as a stepping stone to a permanent position in the larger field of data work, and the organization provides robust support in terms of job search support, such as interview practice and resume review, along with more holistic career advice sessions.

For the Fellows, DataWorks thus functions as both a job, a learning opportunity, and a career center. This is in direct contrast, and in response to, the traditional paradigm of data work, as it manifests in crowdworking sites and services, such as Amazon Mechanical Turk (Amazon MTurk) [61]. Data workers on those platforms are treated as temporary, contract workers, who will likely never meet their bosses, or the requesters whose tasks they complete. As a result, data work facilitated by large platforms is often marked by frustrations for both the workers *and* the requesters who post tasks on those platforms. Workers have no job security or promise of stability, lack all traditional job benefits, and are vulnerable to malicious requesters (see 2.5). Requesters, in turn, are unsure of the identity of workers and question their authenticity and ability to complete tasks satisfactorily. While there are arguments in favor of quasi or wholly anonymous working relationships, such as including people who face discrimination in the traditional workplace, it also disrupts traditional workplace-based protections. For example, the platform then must take over the role of mediator between requesters and workers, but as the requester pays the platform to post their task, there is little business incentive for the platform to scrutinize requesters' behavior.

Given this paradigm, DataWorks is unique as a data services provider. The Fellows are all residents in the metro Atlanta area—many of them have always lived in the city. Most of DataWorks' clients are also Atlanta-based, such as local civic and non-profit organizations, along with academics at Georgia Tech itself. Besides communicating on common telework platforms, such as Microsoft Teams, Zoom, and of course email, many clients are able to meet with the Fellows face-to-face in the DataWorks office space at Georgia Tech.

This makes DataWorks an unusual site for data services, as not only are the workers and requesters (Fellows and clients) known to one another, but they share a hyper-local context of lived experience in the city and often work on datasets related to it in some way.

2.2 Motivating Example: the Children's Book Project

In January 2021, I starting serving as the client lead for project from an academic researcher, a project that took much of the Spring semester to complete. Our task was to annotate hundreds of images pulled from a popular children's book, *Goofus and Gallant*, designed to teach young children about proper social behavior. Two identical characters, the titular Goofus and Gallant, react according to their monikers for given social situations. While Goofus has his feet on his grandmother's couch, Gallant is offering to help her set the table.

The data workers were being asked to annotate these scenes, namely to to describe the behavior that they saw. Their directives were to use "clear, concise sentences" when describing the scenes, but to avoid inferring or otherwise commenting on the apparent mental state (or mood) of the characters. However, describing their facial features was acceptable ("the boy is frowning" would be acceptable, but "the boy is upset" would not). The client intended to use this dataset of annotated images to train a computer vision system to recognize scenes in children's books more broadly, with the Goofus and Gallant series serving as the first step in the future system.

The work took three months, in part, because we (the workers, and myself, as the project lead) couldn't get it right. Workers had two key problems. First, there was disconnect between the standardization the client wanted, and the uniformity workers through they were expressing. For the workers, there was no difference between "He is sad" and "he is SAD," and from a semantic perspective, they are entirely correct. However, the ML model, which is designed to take in a cleaned (standardized) dataset might treat these descriptors as unique tokens, for example, different "sad" from "SAD," or assume a different

QID3.

QID4. The boys were playing with toys.
The boy in the red is washing his hands in the sink.
The boy in the green is looking at the boy in the red.
The boy in green is picking up the toys.
The boy in the red is smiling.
The chair seat pad is green.

Figure 2.1: Example book image, with workers' descriptions below.

semantic relationship based on the capitalization of letters. Workers were also inclined to use vernacular language, as most people naturally would when describing the events of a picture. "He's goin shopping" is, again, semantically reasonable, but, given that the dataset will be fed into a model relying in part on syntactic similarities, that is meaningfully different than "He is going shopping," which was the version the client expected. While I could employ some aids, such as regular expressions, to edit out some variance, it was time-consuming, frustrating work. A few weeks into the project, after several of these editorial sessions, I realized the issue was not that the workers weren't trying to standardize their text, but rather that they didn't know *what* needed to be standardized, because they didn't have a sense of how the model would be employing their descriptors.

The second obstacle was the seeming triviality of the work. Why did it matter if we focused on the "wrong" details (such as describing the color of the couch, when the main focus should be who was sitting on the couch)? What was the difference, really, between describing a frown and a bad mood? Similar to the standardization issue, the short answer had to do with how the model would need to discern information from the text the workers

were generating. If we submitted the wrong text (or focused on the wrong details), the larger model and subsequent system would be unable to decipher what was important in a given picture.

However, there was another element at play: worker agency. Workers were reasonably frustrated by what felt like an endless game of cat and mouse with the seeming ambiguity of the task description ("Describe what's in the picture") and the hyper-specific commentary they would get back ("Don't talk about color of furniture.") At the start of the project, the client and I had emphasized to workers the importance of employing their own perspectives and ideas, highlighting what we saw as creative license. When the workers tried to exercise that license, they were instead met with detailed criticisms from what felt like an invisible rule book.

Together, these obstacles represent a lack of *critical data literacy*, or the ability to understand how a dataset has come into being and could be or has been employed [12]. This is a step beyond regular data literacy, which requires a fluency in reading figures and understanding statements generated based off data [10, 62]. *Critical* data literacy requires all of those skills, and, in addition, being able to reflect on, and critique, how that process took place, and what it might have been motivated by, or whose agenda it fits into. Personal agency is also a fundamental co-concept to critical data literacy; in order to practice critical data literacy, an individual must be able to understand their own work or practice in relationship to a larger data-intensive system and feel empowered to act according to their understanding of that system. The workers were lacking both these tenets of critical data literacy during the project. First, the workers didn't understand what the larger chain of production was—or, what their annotations would be used for—nor how they fit into it. The second feature was the lack of personal agency; the workers felt totally disconnected from the larger goal of the project. Critical data literacy pulls from the Friere's notion of critical consciousness [11], which is to say, without the ability to act on the critical insights, it not always an immediately useful skillset. Said differently, true critical data literacy in

a workplace setting also requires agency, which must be fostered in part as a democratic workplace.

In the children's book project, workers were not yet familiar with the critical data literacy approach, nor did they feel empowered to exert their agency as data workers. Notably, this is not to say that workers were unable to practice critical thinking skills—very much the opposite. The problem was translating the critical perspectives they already knew and applied, often times very acutely, in their everyday lives to the practice of data work, particularly as it pertained to their professional roles. It should be noted that this is a different set of expectations, too, than is usually expected in the hourly wage work contexts from which many of the data workers were coming from[1].

Further, in being unable to envision the inner workings of the model, they were unable to assert their agency, namely in the form of their perspective. What the client and I had touted as a angle of creativity relied on a nuanced understanding of creativity, particularly that specific to algorithmic systems. The space for personality—for lived experiences—as the client saw it, was in the ability to make calls about what was happening in the scene, or how actions were being enacted and received by various characters. The premise of the books also lent what we (both computer scientists by formal training) understood as space for, again, a kind of computational creativity, or a chance to speak back to social behaviors deemed (un)acceptable. For the workers, however, that element of creativity was missing, since they weren't able to envision themselves as part of a larger computational system. Specifically, their role in the work and the calls that they made on behavior seemed limited to the scope of the project, rather than generalized through the computational angle of the project.

[1]These tensions are explored more in work by Johnson et al. [63]

2.3 Identifying What Data Workers' Critical Data Literacy Contributes to Data Work

Not only did I see what a lack of critical data literacy did to workers, in terms of their interest and investment in the project, but I also saw how it impacted the data annotations that we were producing. Unfortunately, this lack of critical data literacy among data workers more broadly is not unusual in the larger data work field. Reflecting on the experience, I saw the need for a critical data literacy crash course for the workers, as an example of what critical data literacy could bring to data work, to the benefit of both workers and the requesters of their labor. Again, workers' lack of critical data literacy skills was not because they didn't know how to be critical or employ a critical perspective. Instead, because they were unfamiliar with the technology being used, specifically ML, they weren't sure how they fit into the larger system design. Further, they had little prior experience making sense of the behavior of computational systems in workplace settings. It is difficult to critique a system that you do not yet understand the workings of.

While DataWorks serves as one example of the need for critical data literacy (CDL), it is by no means a singular one. The role of CDL in data work more broadly is an issue for 1) system correctness, or how well a system performs given its initial training data, and 2) ethical workplace practices, particularly concerning the move towards digitally-based future of work. Said differently, when the requesters and data workers share a mutual understanding of machine behavior, workers are better able to perform the labor of dataset assembly and curation. This benefit for requesters also requires raising workers to the role of collaborator and partner, including a profound reconsideration of the exploitative workplaces and deprofessionalization of data work as a craft.

Beyond DataWorks, larger scale issues of system correctness in data-intensive systems, particularly AI and ML ones, continue to plague these widely-deployed systems. For example, consider ImageNet, a foundational computer vision benchmark dataset debuted in

Figure 2.2: A mislabeled ImageNet entry (ID: 00023975), reproduced from [65]

2009, and one of the first major dataset assembly projects to be deployed on Amazon Mechanical Turk [6]. ImageNet catalyzed a wave of growth in computer vision, via the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), and the academic paper introducing ImageNet has more than 25,000 citations, at the time of writing[2]. However, there are plenty of errors (about 3.4% of the total dataset) in the classification of the images that comprise ImageNet, dutifully recorded in a project shared more than 10 years after ImageNet's creation, in 2021 [64]. Some of the errors are mundane, though it is not a far reach to imagine the unintended consequences, particularly for marginalized or vulnerable populations.

While it is impossible to know exactly why these images were mislabeled, and under what conditions they would have been correctly labeled, the harm incurred is not insignif-

[2]See https://ieeexplore.ieee.org/document/5206848

icant. Fei Fei Li, the ImageNet project lead and a famous figure in computer vision, felt a "twinge of culpability" ([66] as reported in [67]) upon hearing that the Google Photos automatic image recognition system was labeling Black people as gorillas [68]. The very basis of the labels for the 2012 edition of ImageNet (ImageNet-ILSVRC-2012) is *WordNet*, a dataset of English-language words complete with semantic relationships between entries in the dataset [7]. Researchers Prabhu & Birhane describe the deeply problematic, offensive, and NSFW labels contained in *WordNet*, including the N-word and a host of nonconsensual pornographic terms [8, 69]. The same issue plagued the former *80 Million Tiny Images* dataset, which the creators chose to remove from the public sphere entirely, citing the difficulty of verifying the acceptability of the images [70].

While removing the dataset from web (and consequently commercial usage) was likely the right choice, **is it possible we could have foreseen the problems before these datasets became standard choices on which to build all kinds of automated systems?** In particular, if the data workers annotating these datasets had true critical data literacy, both in the skillset and the ability to practice it, without retribution by their employers, could they have alerted system developers of problems with the dataset, halting its use until they were remedied?

A hypothetical defense the dataset developers might offer is that the project was simply two enormous to be subject to rigorous research team review. However, there is a missing entity, here: who was not only reading through many of the labels, but in fact also in charge of applying them to a given image? Those close to the dataset, here, the labeling of datasets based off of *WordNet*, might have been able to warn of potential for misapplied labels, unclear directions, or generally flag other sites for bias and misidentification. To be clear, this is not to in any way point to the labelers who worked on these computer vision systems as holding culpability for application of these labels; instead, we should question why the labelers were not given the agency to point out these concerning labels, or otherwise protest labeling racist and nonconsensual pornographic images. To give opportunity

for data workers to raise red flags or otherwise voice concern also requires treating them as collaborators, namely providing space for such reflection. Given that much of image annotation, and data work more broadly, now occurs on a contract basis, via large platforms like Amazon Mechanical Turk, there is little incentive for workers to engage in such behavior, given that they are not paid for such time and they also risk upsetting a requester who could then block them from future tasks.

What would meaningful employment of data workers look like, as it pertains to system correctness? Kapania et al. explored the perspectives of data work requesters who expressed interested in having diverse data workers perform their tasks, but those requesters failed to design annotation workflows that made use of that diversity [71]. Instead, meaningful employment opportunities would mean both making the space for worker reflections on the state of the dataset—given that workers are, by virtue of being those to develop it, best suited to declare that dataset's contents and limitations. Engaging these perspectives can be done by making workers collaborators and equal project agents, rather than surface attempts to correct dataset imbalances. For example, we can imagine how early warnings of dataset composition and over-saturation of a limited demographic pool would have resulted in different system deployment, perhaps one that did not result in Google's post-hoc attempt to re-balance their facial dataset demographics by sourcing images of Black men from Atlanta's homeless population [72].

System correctness, thus, goes hand in hand with worker agency and rights. When data workers lack agency—in both co-defining the terms of their labor (such as through a union) and negotiating their role in the development of the eventual data-intensive system—the entire system suffers from their lack of input and perspective. Previous work has shown the extensive implications of what happens when this expertise is ignored. For example, Miceli et al. argue that *bias* is not the product of faulty, incomplete, or improper work by data annotators, but rather the "interpretive character" of these tasks, as they are designed by dataset requesters [20]. Sambasivan & Veeraraghavan explore how data workers, specif-

ically field workers, are subject to role minimization in the development of large datasets and, should that understanding of field workers be re-oriented to one of productive partners and collaborators, the larger process of AI development would benefit [73]. In the next chapter, I share my own example of the implications of when data workers are endowed with the agency to voice their opinions and perspectives in the workplace, they are able to correct emergent problems that the dataset requesters may be entirely unaware of.

2.4 Theoretical Lens

My theoretical approach is one focused on the role of understanding and documenting the role that subjectivity plays on datasets. When we understand what datasets contain and how they were developed, we can safely employ them. This understanding of data is, at the highest level, an enaction of Haraway's idea of *situated perspectives* [31], particularly as it pertains to data. The field of critical data studies is no stranger to Haraway's ideas, and in appropriating them myself, I follow in a larger, if nascent, tradition of *human-centered data science* (HCDS) [74, 75]. Chancellor describes human-centered machine learning (HCML), a sibling of HCDS, as an acknowledgment of the myriad negative consequences of machine learning, both intentional and not [76]. For Chancellor, HCML is hard to define in precise terminology, but the focus, Chancellor argues, should be reoriented from singular emphasis on the product, also making sense of the human impact. Kogan et al., returning to discussion of HCDS, the adjunction of "human-centered" to data science is premised on a need to recognize an account for the "highly situated and nuanced nature" of data worlds [75].

Said differently, practitioners of HCDS (or HCML) are those who do not believe that data presents a neutral, incontestable archive of events. Instead, HCDS and HCML acknowledge the social worlds and forces which contribute to give rise to a given dataset. Notably, while current programs of study in HCDS have explored these social factors as they influence data scientists (for example, [77]), an emerging group of scholars has started

to concern themselves with everyone who "touches" a dataset, or in any way contributes to its development. This focus embodies what D'Ignazio & Klein have termed *data feminism,* specifically the application of the feminist lens to data science [47]. A feminist data science practice is one that observes, interrogates, and finally addresses the differential impacts of data science products and technologies, particularly along intersectional lines. Along with a focus on justice, data feminism also represents an embrace of pluralism, particularly as it pertains to lived experience. While a traditional notion of data science often posits data in the objectivist tradition and its close relatives, a pluralist, as feminist, data science requires that we question how and why a dataset has been desired, developed, and potentially shared, and who performed that labor.

Applying the data feminist approach, as a strain of the larger HCDS practice, several scholars have begun to upend the hierarchy of attention. Within the scholarly practice, attention, in the form of ethnography and workplace studies, often results in not only in focus on the subject at hand, but also in the development of new tools, workflows, and grants to help further a practice. As many accounts of how the practice of data science takes place has focused on those at the later stages of the process—namely data scientists, machine learning engineers, data engineers, and so on—a narrative has emerged. This narrative suggests that the problems with data science practice, or the (un)intentional unfair effects of data science products and technologies, are somehow the fault of the labelers, annotators, and other workers who create these datasets. The solution to the problem, we are told, is to figure out how to "unbias" the dataset. While this "debiasing" is impossible from a data feminist approach, as bias is the result of social worlds reflected on a dataset as it is composed and consumed, is it also pointing the finger at the least powerful actors in the larger data science pipeline. As I have argued earlier in this chapter, in the event the individuals working on dataset annotation want to speak up, they are almost uniformly discouraged from doing so.

Here enters the idea of *studying up*, as originally formulated by Laura Nader in the

field of anthropology. To *study up* is to reposition attention on the upper social strata of a given social group, not just the lower strata; in doing so, ethnography becomes additionally a study of "power and responsibility" with the potential for change including those who have the most ability to enact those changes [78]. Miceli et al. make use of studying up to demonstrate how agency is removed from data workers who perform the labor of dataset annotation and labeling, and how so-termed "biased" data science systems are really representative of the choices, directives, and cultural understandings of those who command the creation of those datasets [20]. I follow the call of Miceli et al. in aiming my research lens not only toward the subjects at hand (data laborers) but under what auspices they operate, as designed by dataset curators, such as data scientists.

I push the notion of studying up one step further, however, at it pertains to data work. Much data work occurs on platforms such as Amazon Mechanical Turk, which enables dataset or data work requesters to distribute that labor as discrete micro-tasks, as a practice that has become known as *digital piecework* [23]. There are two key co-contaminate problems with platform-based data work: 1) workers, as contractors, are subject to the whims of requesters and lack fundamental workplace protections, and, 2) requesters are generally frustrated with the work they must do to verify the identity and submission quality of worker submissions. When we examine the *triangular* shape [61] of stakeholders in platform-based data work, where all stakeholders (workers, requesters, and platforms) have a link to one another, the requesters emerge as the most powerful party. Prior work on the Amazon Mechanical Turk platform, namely in the form the mutual-aid TurkOpticon project [79], attempted to push responsibility firmly to the platform's domain, given that they exert a great deal of control (as a mediating party) between requesters and workers. The goal of the project was to unite workers with the aim of forcing Amazon to enact wide scale change. However, Irani & Silberman, in reflecting on the success of the TurkOpticon project—which was widely adopted by Turkers (workers on Amazon Mechanical Turk)—saw it as ultimately a failure, in the sense that it did not force Amazon to enact the desired

changes to the Mechanical Turk platform [80]. The Turkopticon project is in many senses a success; the project is still up and running and continues to be heavily utilized by Turkers. However, Turkers bear the brunt of the labor in providing reviews and ratings of requesters and their tasks, while Amazon merely reaps the reward of a consistent worker population.

Employing the idea *studying up* to the triangular shape of platform-based data work, I isolate a stakeholder group—requesters—who have rarely been identified as potential agents of change, despite the fact that they hold outsized power and responsibility within the ecosystem, as the ultimate employers of the data workers [81]. While prior attempts have pointed to a limited set of factors requesters should consider when creating a task to post on a digital pieceworking platform. For example, requesters have been asked to consider a $15 hourly wage for workers [82]

However, these efforts lack an overall view of the requesters as a group with power to determine not just the terms of labor, but the experience of work. Review work, such as Hawkins & Mittelstadt's expansive review of AI data enrichment practices [83] or Scheuerman et al.'s examination of computer vision dataset curation practices [84] , have found that the common contemporary requester practices go under-reported, if even ethically configured to begin with.

2.5 The Status Quo in Data Work

Given that there are multiple strains of data work, in this section I address the overarching concerns for the lack of worker perspective inclusion in data work, as it pertain to generalizable system development (namely AI and ML systems). I then focus on the most prevalent model of data work, one in which a platform mediates the task delivery and completion relationships between workers and requesters.

2.5.1 In data work at large

Perhaps no other category of work is more representative of the potential, alarming future of work than data work. As I expand on in the following chapter, the history of data science is one that is intrinsically tied to the exclusive West Coast libertarian identity of the 1960s and 70s. As a practice focused on generalized, global connections, data science has a history of glazing over the nuance of a given dataset—namely, its sociotechnical origins and depictions of populations, as scholars like D'Ignazio & Klein [47] and Loukissas [85] have described. This emphasis on generalization, naemly building a system that can be deployed to any situation or context given enough data, is an underlying assumption of data science. Perhaps the most extreme form of generalization is that applied to individuals and lived experiences; these systems are modern incantations of the pseudosciences of physiognomy and phrenology ("Physiognomic AI," as termed by Stark & Hutson [86]) and target already-marginalized members of society, presenting the opportunity for widespread discrimination.

The ethos of these overly-generalizable systems is clearly not only in deployment, but also the development of these systems. Data work is required to facilitate the datasets needed to train these systems, but that data work is shaped by the priorities of dataset requesters, regardless of the realities the data workers performing that labor encounter as they go about it. Consider object-recognition systems, such as those trained on datasets like ImageNet. DeVries et al. found that these systems, meant to identify a given object in context, fail to recognize common household items in countries with "low household income" [87]. The researchers theorize that the lack of recognition of these objects, as photographed in low-income countries, demonstrates in part that these objects occurring out of expected context (as in high-income country, e.g., soap dishes only appearing in the bathroom). While misidentifying a soap dish might not be a an immediately serious concern, we should be concerned with that these inabilities to recognize and respect the context of low-income countries means, and what generalizations will be made about populations

in those countries as a result.

Now, much of the global data annotation and labeling workforce is located in relatively low-income countries [22, 88, 89]. So why don't these systems represent a more diverse (or culturally expansive) perspective? In general, the voices of data workers (as data collectors, annotators, labelers, standardizers, and organizers) are silenced through a series of overt directives and system design features, despite representing a broad range of lived experiences [90]. Whether the work takes place at a subcontractor (or business-to-business firm) or on a digital task platform (such as Amazon MTurk), data workers are disincentivized from speaking up [15, 13] and when their perspectives are allegedly valued, they are not put into practice [91, 92, 71]. This is consistent with studies of worker creativity on crowdsourcing platform, which have found that requesters lack good practices on harnessing crowdworker creativity [93].

The minimization—or even effective elimination—of data workers' perspectives is indicative of their treatment in the larger data ecosystem, too. Once the data labor is extractor from workers, for frequently sub-standard wages [94], workers are left with nothing to show for their efforts. Data workers generally do not benefit from the systems they helped to bring into being, nor are they paid dividends on the eventual profits generated by those systems [95]. This is consistent not only with trends in gig work more broadly [88], but also contemporary Taylorism [96].

Nor is most of data work transferable as a demonstrable skill. On digital task platforms like MTurk, workers must constantly re-qualify for tasks to prove themselves capable of the tasks they wish to undertake, which is frustrating, given that it wastes time for which workers are usually not being compensated [97, 98]. While there are several efforts to continue to develop worker skill level, e.g., [99], these initiatives again place the burden for professional development on individual workers, as opposed to their employer—which, as workers are technically classified as contractors (as opposed to traditional employees) on most platforms, effectively positions task requesters as workers' employers.

2.5.2 On quasi-anonymous task platforms

The workers-as-contractors positioning is important, because it also leaves individual workers responsible for their skill-building and welfare, in a labor ecosystem that is notoriously harsh on both. Several initiatives have been put in place besides the TurkOpticon project [79]. However, these projects have been relatively unsuccessful in rebalancing the overarching dynamics that still heavily favor requesters [61]. While many of these sites would also be considered crowdworking platforms, not all crowdworking platforms are used for data tasks. For more creative tasks, where recruiting is based on a portfolio or work samples, the worker's identity is usually known to the task requesters—for example, UpWork.

Platforms like UpWork usually follow a traditional freelancing model, merely digitizing the task marketplace [100, 101]. In contrast, data tasks are usually posted on quasi-anonymous sites, where the identity of the worker is not known to the task requester, and the duration of the working relationship is usually shorter; for example, the time taken to complete an individual task can be no more than a minute or two, while on non-anonymous platforms, they are usually longer term projects, on the order of several hours or even weeks, and sometimes months. Said differently, the opportunities a worker has to demonstrate their skillset, as freelancers normally would [102], are extremely limited. Howcroft & Bergvall-Kåreborn further distinguish between the two kinds of platform positions, describing the quasi-anonymous model as "online task crowdwork," where each worker is a generalist and the atomization of work fits into a Taylorisation model, while non-anonymous sites are "profession-based freelance crowdwork" in which the labor is typically more specialized and requires substantial professional experience and knowledge [103]. Thus, besides with the anonymity comes also an emphasis on the replaceability of any individual worker.

It is on these quasi-anonymous sites that task requesters needing the work of a large numbers of workers usually post their tasks. In this dissertation, I focus on these quasi-anonymous platforms, in which workers are seen as expendable and exchangeable for one another. To the requester, individual workers are usually no more than an alpha-numeric

string of characters (as a worker ID) which is the only distinguishing feature between workers; while there are known flaws in this privacy model (such as on Amazon MTurk [104]), it is a paradigm that does not immediately identify or personalize workers. As these platforms are best suited for large-scale projects—consider that Amazon MTurk made the ImageNet project possible [6, 105]—they play an important role in the development of datasets for AI and ML systems.

A sampling of these issues are as follows:

- *Privacy and security.* While data workers on quasi-anonymous digital data work platforms are theoretically identified only by an alphanumeric string that can be used to track a given worker between tasks. Setting aside the fact that workers can still be de-anonymized with requester-designed surveillance tactics [104], even the experience of work sets workers up for security and privacy vulnerabilities. Given the imbalanced power relationship between workers and requesters, requesters are able to take advantage of platform workers, in terms of violating their privacy [106, 107]. Namely, requesters—even if they are barred from doing so by platform rules—not infrequently ask for workers' personal information, and workers feel compelled to give it, since they are reliant on the requester to approve their task submission. Further, workers feel compelled to give up their personal information out of fear for losing access to further work and because they see it as a contribution to scientific research [108].

- *Accessibility.* There are two components of accessibility. First is physical accessibility, such that workers are both physically able to engage with the platform and have the materiality to do so. Uzor et al. found that many standard features of digital data work platform tasks, such as time limits on task completion or visual presentation of the task itself, make it difficult for users with physical disabilities to work on these platforms [109]. Various alternative platforms have been designed, including voice-based ones for workers with visual impairment [110] and ones that work on mobile

devices, seeing as many crowdworkers do not have access to high performance computers [111]. However, alternative platforms have had far less traction, and the market is still dominated by a few major data work platforms. Further, common work practices of crowdworkers on these platforms are not supported by requesters—for example, working on multiple tasks at once [112]. Similarly, the supposedly flexible hours of crowdworking are actually much more inflexible, since workers still must adhere to the schedule of requesters, in order to get access to the limited number of tasks available [113].

Second, accessibility as equal opportunity between social groups, is a concern on digital data work platforms. Beytía & Wagner find that micro-task platforms, including digital data work platforms, actually exacerbate existing gender gaps [114]. Among other concerns, the promise of micro-work platforms as a tool for building digital skills does not hold up—instead, women are again relegated to lower-prestige computing work, as has occurred repeatedly throughout recent history [115]. Despite the fact that women were at the forefront of computing, they were pushed aside as soon as computing became a desirable profession [116, 117, 118, 119]. Further, several of the attempts to counteract other issues in this list, such as worker-sourced ratings of other workers, are again sites of gender discrimination, e.g., [120].

- *Collaboration.* While traditional workplaces support knowledge sharing based on shared workplace proximity, workers on digital data work platforms are not able to benefit from this [121, 82]. Lack of collaboration has implications for things like organizing and fighting for fair pay, but this lack of shared experience also makes it more difficult to build skill as a data worker [122, 123]. While worker co-ops have been proposed as one option [124], other operations where workers join together have been termed as violating the terms of service, of, for example, Amazon MTurk [22].

- *Arbitrary and malicious judgments.* Given the structure of these platforms, many workers have numerous employers (or work for multiple requesters) during their working hours. This can make it difficult to keep track of which task was actually performed for a given requester. Compounded by the fact that the default setting on these platforms–from the requester side–is to accept or reject a task without giving any explanation, workers rarely have insight into why their task may have been rejected. This lack of transparency is such a problem, for example, on Amazon MTurk, that the Turkopticon workers group has launched a campaign to raise awareness of Mass Rejections—, or when a requester rejects all task submissions (at no penalty to the requester) and their account funds are automatically reloaded to starting balance, allowing the requester to receive a second (or, n-th) batch of fresh work for no additional cost [125]. This has direct harmful implications for workers; first, they may lose standing on the platform (e.g., having their prior task acceptance rating dip) which can reduce the number of tasks they are qualified for[3]. Further, in order to get that prior task acceptance rating back up, they may have to work on tasks that are lowly paid (even by crowdworking standards) or outright malicious (e.g., require disclosing personal information or otherwise against platform Terms of Service) [126]. The lack of decision transparency, regarding submission acceptance or rejection, is also sometimes the result of algorithmic management methods, which have been found to correlate with a lack of meaningful work experience [127].

- *Trauma.* Given the repetitive nature of much data work, when the work itself depicts violence or disturbing content, the experience of viewing that material is compounded by the sheer number of tasks. The experience of content moderation workers, as one group of data workers, gave an early indication of the potential trauma performing this kind of labor can induce [89, 128]. More recently, Kenyan workers are suing Meta and their subcontractor, Sama, a self-termed ethical AI company, that

[3] See a discussion between Turkers on the r/mturk forum page: here (archival link: here).

provides data labeling services for major tech companies, under the mission statement that "living wage work is the best way to lift people out of poverty." [129] [130]. The Kenyan Sama workers spent their working hours watching video of "murders, rapes, suicides, and child sexual abuse" [131] that had been flagged by Facebook (now Meta). Sama is facing similar allegations from a second group of Kenyan workers, this time for work Sama took on from OpenAI, to refine ChatGPT [132]. In this case, besides a host of more classical labor violations (e.g., arbitrary dismissal), the workers describe being exposed daily to endless streams of disturbing content, leaving at least one moderator with significant emotional disturbance, along with destroying his relationship with his family [133]. There are also recent reports that platforms, e.g., Tokola [134], have a number of underage children among the worker population, on the platform designed for adult workers.

- *Lack of professional development.* While population of data workers on platform sites is large and in constant fluctuation, making it hard to get definitive statistics about the population, we do know that a number of workers on the platforms are doing so as a full time job, e.g., on Amazon MTurk [135]. In some cases, workers are stuck on the platform, even though they hope to move off crowdworking sites and transition to more traditional stable jobs, seeing platform data work as a temporary solution [136]. In other cases, workers are committed to the idea of data work as a career, making things like professional development all the more important [91]. At the same time, workers are losing their personal identity as a result of their employment on crowdworking sites [137]. In some cases, they develop a professional identity that is distinct from their personal identity [138], which again troubles the idea that working on such platforms is merely a way to make a couple of dollars and kill spare time.

All these difficulties are presented in conjunction with the generally abysmal pay that most digital data work platform workers receive [139, 140]. In general, the economic and

labor situation of the data workers performing the data labor that makes possible AI and ML systems is far less prestigious and enjoyable than that of the system developers, e.g. tech company data scientists or developers [13]. This differential treatment is consistent both with the historical precedent set by developers-as-prestigious, once computing became a hot area [116, 141, 142], and with what other researchers in this space have observed. Widder & Wong describe this as one site for ethical investigation, as part of a larger set of ethics and policy opportunities presented by the neglected early stages of the AI supply chain [143]. Specifically, dataset creation and curation, as data work, should be seen as modular steps within the larger AI supply chain [144], but rather as overarching concerns with implications for system development and use.

In response to the treatment of data workers on digital task platforms, among other concerns about the development and deployment of AI and ML systems, there have been numerous calls to reform our adoption and uptake of AI. Resistance occurs on two levels. First, the microlevel—or close to the data itself—where individual datasets and data tasks are subject to refusal, on the grounds of what they might contain or what they may be used for [145]. Second, AI itself is contested, as a macrolevel of resistance; consider, for example, the calls McQuillan makes in "Resisting AI," that AI is fundamentally dehumanizing. Bender makes a similar argument, explicitly including ghost work (as representative of AI and ML data work) as an example of the dehumanizing properties of AI, focusing on the quasi-anonymous model that encourages distance between requesters and their workers. My own work is premised along the same lines, focusing on the way that humanity can be brought to AI, or fundamentally repositioning AI as byproduct of lived experience and perspective.

These movements fit more broadly within the theory of feminist refusal [146]. Beyond issues relating to data workers, feminist refusal has also been seen as a way to resist datafication efforts more broadly; for example, the feminist data manifest-no [147]. While I am intrigued by these projects, my work is not one of direct resistant to AI. This is because,

whether I like it or not, AI is not going anywhere in popular commercial and governmental use. Given that we are faced with these systems, my work is premised on making them the least harmful versions they can be.

However, in performing this work—or in getting data workers recognized as skilled early auditors of datasets, which requires an understanding of possible future uses of that dataset in AI systems—I do work to get more eyes on a given dataset and sequential AI technologies. This opens the door for more discrete kinds of resistance, such as whistle blowing and *soft resistance* [148], as well as more direct data refusal [145]. Depending on the individual practitioner's work status and ability to withstand retaliation, which, for many data workers is extremely low, given their contractor status and need for income [149, 150], these actions may be appropriate.

2.6 Research Focus

In thinking about how to return the human perspective to AI, we can turn to Charles Mills' concept of the non-ideal, or the idea that the distinguishing characteristics of the non-ideal can help us understand that the unspoken ideal is [151]. Applied to data work, "human computation" and the idea of ghost work, as invisible labor, suggests that the ideal is a system that requires no human actors for operation. In fact, obscuring any human role in computation plays into fears about the future of AI, namely that it will be intellectually superior to humans, or the so-called "singularity" [152, 153]. **Reanimating, or un-obscuring, the role that humans play in development of AI and ML systems at all levels, but especially in data collection and curation practices, reminds us 1) that these systems should be designed as supplements to, but not replacements for, human intelligence, intellect, and lived experience, and 2) that it is those very human characteristics that shape our systems.**

In particular, our current class of generalizable systems (namely, AI and ML) are plagued by a host of issues which are popularly known as *bias*. As a term, **bias** is tricky

for AI practitioners to grasp, particularly when it comes to ethics; instead, as Browne et al. found, the term is not clear to AI practitioners, nor does it flag practitioners are something they are directly responsible for considering [24]. Further, when we think about bias, as Miceli et al. argue, we are quick to point at the subjective experience of data workers as bringing an unwelcome bias into an otherwise value-neutral data work ecosystem [20]. Instead, we are discussing *power* in multiple dimensions—who commands the creation of a dataset, for what ends, and who is represented in that dataset—and to label these power imbalances as bias overlooks the core power differential that take place, namely those between system designers and data laborers [51].

In fact, instead of ad-hoc (or post initial system deployment) solutions to the discriminatory behaviors we see in AI and ML systems, such as re-balancing dataset composition, re-tuning system weights, or even pulling back system deployments, what if we built systems correctly in the first place? Specifically, what if we broadened the perspectives that are represented by our generalizable systems? Critically, the inclusion of broader perspectives should not become the "participation washing" that Sloane et al. observe in experiments with participation in machine learning [154]. Rather than inviting outside perspective only once a system has been designed, these generalizable systems should be built with a multiplicity of perspectives from the ground up. Further, the garnering of these perspectives should not take the form of "prospecting" as are common in data science [155]—**facilitation of the integration and acknowledgment of these perspectives should be genuine and occur from the earliest stage of system development.**

My contribution to the profession of data work is the demonstration and deployment of tools to situate and deploy data workers as dataset auditors, to the benefit of everyone in the AI pipeline, from developers to end users and bystanders [48]. First, through DataWorks, as an alternative work site predicated on creating lucrative and sustainable careers for mid-skill data workers, when data workers are trained in practicing critical data literacy, they *are* able to serve as early auditors of the datasets, in both material and conceptual di-

mensions. I demonstrate this through the AAVE project case study described in Chapter 4. Second, I contribute a range of designs (including cover sheets 5.1, the Datum Fieldnotes tool 4.3) and insights about workplace dynamics, from the troublesome platform-mediated ones (5.2) to the promising ones of civic and non-profit data professionals (3.2), needed to understand how to scale up the opportunities for dataset auditing by data workers at sites beyond DataWorks.

The sum of these contributions is the demonstration that our current paradigm of data work undervalues the professionals who practice it. By both increasing our respect of the practice and its practitioners, we stand to gain a massive amount in terms of the overall safety and reliability of AI systems. Further, it gives us the opportunity to develop systems that better reflect a broader understanding of concepts in the wild (think of the multiplicity of something as mundane as stop-signs globally) vs the limited experience of uniquely that of AI designers and developers.

Positioning data workers as having the potential to become dataset auditors should not be understood as a suggestion that the partial perspectives of individual data work professionals will produce an objective evaluation of a dataset's values. There exists no objective record or reflection of dataset's description of people or events, because there exists no one universal truth (or impression) of such things in the wild [156]. While the classification impetus of AI systems demands clear discrimination between components, the reality of those systems, as socially embedded ones which equally reflect social conditions and proceedings, is discordant [157, 158].

Instead, in scaling up the idea of the data worker-as-auditor beyond DataWorks, I hope that this work will be understood as reflecting the idea of *cosmolocal scalability* [159]. In the cosmolocal model, scalability diverges from the "at all costs" model to instead prioritize multiple ways of knowing and living. Unlike traditional capitalist models, the cosmolocal scalability approach is characterized by appreciation for diversity, local adaptation, and the sharing of knowledge [159, 160]. While initial models of cosmolocalism hew closely to

the sustainability of production, with an ecological focus, there is an increasingly the turn towards digital communication networks [160]. The emphasis here is on the infrastructures that can be used to facilitate the cosmolocal approach.

The creation of global connectivity, or of broadly applied AI systems, is indivisible from localized practice. The systems produced thus blend a more global outlook with the realities of local contextualization; in practice, this is produced by localized accounts of different dataset auditors reflecting a view of events and people that is more globally minded than that of a small group of AI system designers and developers.

Further, in trying to transfer the learnings of critical data literacy and worker empowerment at DataWorks, and the subsequent implications for dataset creation and development, the cosmolocal approach reminds us never to treat data production sites as interchangeable. While the pros and cons for individual sites and platforms, for the professionals who are employed there, may be better suited to one worker's goals and experience vs a different worker's, the goal here is not to dismantle all sites in favor of the DataWorks model. Instead, the contribution to the body of knowledge about data labor and professionals' experiences and capacities should be applied, as appropriate, to other sites of production. What makes a democratic workplace at DataWorks will not necessarily develop similar avenues of worker advocacy regarding dataset contents at another site. Each site of dataset product and labor is unique in its sociotechnical and cultural context(s).

This dissertation contributes a novel angle on the professional purview and capacities of data workers, specifically honing in on the ability of data workers to act as dataset auditors. In sharing the environmental and social factors that contributed to this development, I highlight the ways in which it was developed specifically at DataWorks in the first half of this dissertation. In the second half, as I examine opportunity for the capacities and mechanisms needed to transfer and apply these learnings more broadly.

However, in studying these opportunities for transfer, I to hew closely to the notion of cosmolocal approach. This is why I stop short of actually testing the deployment of the data

worker-as-auditor mechanisms on sites beyond DataWorks. To fully appreciate the uniqueness of each site would require a through study of the relevant sociotechnical dynamics, as well as partnering with stakeholders, especially the relevant data worker population(s). While I have not conducted this work yet, it is the next stage of work for which the respective studies of capacity and potential for adaption I have mapped in the second half of this dissertation.

CHAPTER 3

UNDERSTANDING HOW DATA WORK HAPPENS

As a field, HCDE[1] is premised on individuals shaping the datasets they work with [74]. Data science gives a new face to *cybernetics*, or the post World War II study of data-focus automated command-and-control systems. AI, in turn, is a new face for data science. In this chapter, I focus on the development of the field of data science, tracing its legacies to cybernetics, and argue that AI is merely the next phase; exploring data science's failings to account for perspective, I conjecture, helps us understand why similar behavior in the field of AI has led us to where we are today.

So long as data continues to be seen as a reference point for organizational behavior and decision-making, understanding how data is transformed into insight is critical. However, data science as a field allots respect and power to the data science practitioners, as opposed to all the individuals who make possible data-intensive systems. There are a wide variety of individuals performing data work, particularly outside of major tech companies, that do not self-describe as data scientists; instead, they are domain experts who practice data work to obtain insight in their domain of expertise. Unlike data scientists who are focusing on assembling massive datasets for huge, generalizable systems, data workers on a local or domain site focus on smaller datasets, which allow them to engage more deeply with a dataset's constituent datum (or, individual dataset entries).

In this chapter, I focus first on why the field of data science seems to have eclipsed all other kinds of data work and data-derived insight. I trace the histories of data science back to an exclusive Libertarian-informed vein of the 1960s Counterculture movement, or one that saw technology as a way to break down bureaucratic systems. Reflecting on this background, I explore how the contemporary data science practice focuses on gener-

[1] Human-centered data science

alization and lacks nuance related to application in specific contexts. Further, I trace why this valorization of generalizability, or universalism—as a product of data science's early influences—has led to concerning results as non-contextualized data-intensive systems are inflicted on local contexts. This has to due with the origins of data science in cybernetics. Here, I point to *cyberfeminism* as an alternative to cybernetics, positing that the pluralist, entangled approach can take us somewhere far more supportive of human health and happiness. The cyberfeminist approach can not only handle the concept of individual and group perspective, but can guide us towards the design of an alternative data-intensive schema.

Moving from the history of data science into contemporary practice, I examine the data practices of civic data workers. In this empirical interview study, we find that the way cvici data workers handle data is reflective of their own understandings of that data, which they have gone to scrupulous length to develop, through collaboration with partners and individuals of all kinds. This style of dataset contextualization development serves as a model for what an alternative to our current paradigm of dataset development can look like. In particular, those dataset documentation efforts undertaken by the civic and non-profit data workers who participated in this study, gives us a high-level model for what the role of data workers as data auditors might look like.

3.1 A Condensed History of AI & Data Science

Many critiques of artificial intelligence (AI) systems are indistinguishable from those leveled at data science (DS). Fundamentally, both technologies are designed to (unintentionally) reduce nuance and context: elements of the human experience, of individual stories and perspectives, collapse into quantifiable data points. Stories of uniqueness are dismissed as outliers; DS and AI infrastructure require a regime of classification [161] that represses variety across the board. This is not out of malicious intent on behalf of practitioners. Rather, reduction happens to diminish time and effort spent on repetitive tasks, in order to synthesize large amounts of information into summary statistics or classifications. The

infrastructure of these systems is statistics, not empathy, care, and understanding. Subsequently, when we use these systems in conjunction with decision making and calibration of human health and happiness, the results are predictably bleak, if not outright destructive.

DS and AI can both be traced to an earlier computational study of command-and-control, the now-passé *cybernetics*. A product of the post-WW2 zeitgeist for automated and unmanned systems of weaponry [162], cybernetics would soon progress beyond its military origins, being appropriated for business and governance purposes as *management cybernetics* [163]. While the current AI wave (preceded by the decades-long *AI winter*, or lack of funding and general interest [164]) is new enough there is little formal epistemology, DS is more mature. Hence, we dig into the epistemology of DS and how its origins in a specific strain of cybernetics shapes its products—the early signs about the epistemology of AI systems, we argue, follow in the same vein. Indeed, the difference between DS and AI systems is often trivial: both "see" relationships between entities as statistical properties, not as humanistic, sentimental concepts.

Consider facial recognition systems' initial failures to recognize faces of color, particularly those with more feminine features [16], which, combined with poor use and understanding by law enforcement, contributed to the arrest of innocent individuals [17]. The problem with these statistical bases is that systems designed with them in mind frequently fail to properly extrapolate, with disastrous effects for human health and happiness.

Facial recognition systems also exemplify the overlap between DS and AI: DS is a study of patterns—or analytical insights—where AI attempts to act on those patterns, creating some product from them. Effectively, DS is an example of *second-order cybernetics* (C^2), which includes the role of the human observer who participates in the construction of models of systems [165]. Our position, as authors, is *third-order cybernetics* (C^3), in which C^2 systems become self-developing (multi-subject and guided by reflexive control) [166].

In studying the epistemology of DS and AI, and their cybernetic origins, we take a step back from the critiques of individual DS and AI systems. Individual algorithmic critiques,

such as audits [167], comprise important parts of algorithmic resistance. However, our arguments should also be aimed at the premises of these systems, or the foundational ideas upon which DS and AI practice has been developed. If we focus exclusively on designing fairer systems, we might miss how computational systems distribute power [47]. Enter here what Susanna Paasonen identifies as *one* of the premises of *cyberfeminism*—or feminism as a critical analysis of cybernetics, which "interrogates and intervenes in technoculture." [39, p. 63] Within this strain of cyberfeminism, we see Donna Haraway's notion of the *cyborg*, a cybernetic organism and a creature of both social reality and fiction [168]. The cyborg opens the door to questioning the role of individuals in cybernetic systems, which we see emerge in C^2 and C^3.

In exploring the epistemic cultures [169] of DS (C^2) and AI (C^3), current literature misses the legacy of cybernetics' universalism as a fundamental epistemological impulse. Starting with DS, Desai et al. and Pietsch's formulations of DS' epistemology are robust but incomplete: DS functions as abstraction, or designing systems that generalize over multiple contexts [170]. This is universalism in practice—DS eschews context in favor of higher level patterns. A similar finding pertains to the epistemology of AI: McCarthy explains how the epistemology of AI is founded on extrapolation of fact, per a series of rules are needed to permit legitimate conclusion to be drawn from those facts [171].

These epistemologies are consistent with what we know about cybernetics, which Norbert Wiener, in introducing the field to the scientific community, described it as a field in which "information is a measurable quantity, and that it can only be studied on a statistical basis" [162, p. 2]. Given new information (or feedback), cybernetic systems perform revisions, begetting further information; cycles are continuous and constantly readjusting. In our era of "Big Data," DS and AI's cybernetic foundations are only increasingly relevant. All orders of cybernetic systems should prioritize correcting system states after receiving negative feedback. However the particular strain of cybernetics on which DS and AI are based is not one that encourages thoughtful application consideration on behalf of its prac-

titioners. The strain is the US incantation of cybernetics by the "morbidly hypermale" [172, p. 138] technolibertarian community; Rosi Braidotti argues that the universalistic stand is associated with men, following Hartsock's notion of "abstract masculinity" [40] [173, p. 354–355]. Resulting systems effectively serve as examples of "informatics of domination" [174, p. 28].

In this work, as authors, we ask what epistemic cultures of DS and AI might look like, if instead based on cyberfeminism. Where technolibertarian cybernetics are the embodiment of the universalist ideal, cyberfeminism – the cyborg—is attuned to local, situated context. This, we content, is the humility that Paulina Borsook dreams of ("Wouldn't it be great if it were programmer humility and awe at the complexity of all things that led to technolibertarianism?" [172, p. 214]).

To illustrate why this imagining is necessary, we begin this article reviewing contemporary DS and AI practice, using Facebook's content moderation practices in the Tigray region of Ethiopia to demonstrate the failures of technolibertarian cybernetic practice (in C^2 and C^3 form). To understand the development and legacy of US technolibertarian cybernetic ideological architecture, we turn to the US West Coast in the 1970s and 80s. Finally, we return to professional humility as one technique to counter universalism, this cybernetic strain's most dangerous legacy, which manifests as ignorance of the nuance of local context.

3.1.1 Situating data science and AI

In 2009, Hal Varian, then Google Chief Economist, termed statistics "the sex[iest] job in next 10 years" [175]. By 2012, the Harvard Business Review argued it was "data science" (DS) not statistics. In more recent years, various internet forums play host to hot debate about whether or not AI will replace DS (e.g. Reddit[2]). The ideologies of cybernetics, DS, and AI are inseparable, but spending some time on what exactly defines these fields can bolster how we understand their epistemologies. As the line between DS and AI is blurry,

we focus primarily on the role of DS, which has a more established epistemology. AI can be understood as an extension of DS, in which the human observer (e.g., data scientist) is instrumental in contributing to generative AI products.

DS exploded as a field in no small part due to the role of data in contemporary business practices. From the 1960s, computer memory capacity has become more physically workable (shrinking component size) and cheaper, consistent with the oft-quoted Moore's Law, which dictates that the number of components per integrated circuit will double roughly every two years [176]. Coupled with more recent innovations in cloud computing (locating data infrastructure at remote data centers, rather than on desktops), we are located in the "Big Data" phase of the anthropocene [177], which Cassel & Topi delineate as the point when data is collected *without* any specific intent [178].

To make meaning from data, the contemporary discipline of DS emerged. "Data science" itself is commonly attributed to William S. Cleveland of Bell Labs, who highlighted it as an area of interested for "government research labs and corporate research organizations" [179]—a series of academic journals were then established: the Data Science Journal (April 2002) and the Journal of Data Science (January 2003). Because DS requires a combination of skills originating in other disciplines (mathematics, statistics, computing) the balance of those skills is subject to debate [180, 181]. C. F. Jeff Wu argued that statisticians should rename themselves data scientists [182], while Peter Naur defined "datalogy" (a precursor to modern DS) as an independent field. Naur called *datalogy* "[t]he science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences" [183]. More recently, Cuadrado-Gallego & Demchenko attempted to create a common "passport" of technical computational skills for DS, to address widespread confusion [184].

Desai et al.'s review of DS epistemologies highlights the primary mode of inference in the field, namely: mathematical and logical reasoning through which deduction, induction, and abduction occur [170]. These ways of knowing also apply cleanly to AI. Pietsch's

critique is more exacting, honing in on DS as an inductivist model of study, opposed to the hypothetico-deductive model followed by most contemporary programs of scientific inquiry [185]. In practice, this inductivist orientation means that DS (and AI) deals in extrapolation; what works in one circumstance is applied to other, similar ones, given sufficient training data.

Regardless of the arguments around terminology and skills, one fact remains: DS and AI are fields preoccupied with generalization over time, space, cultures, languages, and lived experiences. Under the *datafication* paradigm, where data-intensive systems and analyses are afforded ontological and epistemological (faulty) claims of objectivity, audiences fall into the trap of objectivity of quantification [186]. Anything is condensable into data points, given enough of them; this results in a general practice of *dataism*, or trust in data-derived systems and products [186]. With the right dataset, historically nuanced topics become mere fodder for technosolutionist programs and systems.

Perhaps the most commonly lodged complaint by data scientists or AI practitioners is that system failures are due to "bad data[187]," where "bad" corresponds to unbalanced samples, inaccurate labels, or insufficient quantity. Both fields have also frequently claimed "bad" or "biased" data is at fault for en-actions of discrimination across any number of social axes. Unlike data, the generalizable systems (DS & AI) are rendered mere infrastructure, or apolitical tools. This framing neglects the origins and consequences of epistemological descent from cybernetics, or the conception that generalization is a positive practice.

In building generalizable systems, data scientists strip nuance and meaning from the material they deal with. This might be acceptable if we were thoughtful about where and how we employed such systems, but if the recent onslaught of news stories describing harm to humans caused by poor employment of such systems gives any indication, we have been overzealous in our embrace of generalizable systems.

Thus, we trace the origins of DS & AI to identify where practitioners and proponents of

the fields arrived at a consensus that generalization was acceptable and that the nuance of lived experience, culture, and perspective were both quantifiable and of secondary importance. Our conclusion—that a contemporary cybernetics-informed practice still too reliant on the norms established in their origin context (cybernetics)—is coupled with a call for a new kind of data practice: one that valorizes humility and eschews generalization. The problem is not that the data is not *good enough*, but that the systems we are building are fundamentally unfit to generalize over the contexts in which we deploy them.

3.1.2 The Facebook case study

Content moderation is one of the most exemplary combinations of DS & AI systems. Often powered by a mix of algorithmic and human-powered methods, content moderation is responsible for protecting the health and happiness of users on a social platform, and simultaneously makes the platform-operating organization the arbiter of acceptability (sometimes even legality). Facebook (now Meta) is a conservative platform with respect to content moderation practices [188, 89, 189] and quickly removes offending content, largely to avoid alienating advertisers. Facebook employs a classical command-and-control system that features both elements of DS (namely, analytic insights, used in reports and policy making) and AI (for automated removal and flagging of content).

Mark Zuckerberg has stated before Congress that Facebook is not "a media company" [190]. Maintaining that Facebook is a P2P service, Zuckerberg abdicates Facebook's responsibility to vet and monitor content in the way a publishing company is obligated to. In doing so, he infers that tech companies (and the products they offer) exist in a category of their own, as a tool rather than organizations, beyond regular commerce organizations; the point is not how Facebook is used by its customers, but how Facebook understands (or self-determines) its own role. That disconnect suggests that, in Zuckerberg's view, while Facebook has global reach, it only has the responsibility of a piece of machinery.

What complicates these ideas is that Facebook was long run by Zuckerberg's diktat to

"move fast and break things" [191], suggesting that the ramifications of Facebook's product are second to the technical innovation they represent. Zuckerberg's statement does not only let Facebook off the hook when it comes to maleficence taking place via the platform, but it demonstrates a noble goal (global connectivity) without considering the implications of such. Here we see the underlying libertarian cybernetic agenda: simultaneously promoting a non-hierarchical vision of global connectivity, while placing a small Silicon Valley elite at the helm.

In reducing connectivity to interaction between two individuals, Facebook ignores the context of the interaction taking place. The point of a generalizable system—specifically a DS one—is that it works in all contexts, just given the right training data. If Facebook is designed for P2P communication (where P might be not only a peer, but an organization, community, congregation, government, or military entity), then Facebook needs truly tremendous amounts of content moderation resources to cover all of these groups. The cybernetic system's sensors must be calibrated, with a terrifically vast mix of training data (for automated means) and moderator training (for human moderators). Unfortunately, Facebook's attempts have been paltry at best.

Take the example of Facebook failing to engage in sufficient content moderation against hate speech in the ongoing conflict in Ethiopia's Tigray Zone [192, 193, 194] – which the Amnesty International and the Human Rights Watch describe as a site of ethnic cleansing [195]. Facebook posts have incited violence within the larger conflict [196, 197]. From "The Facebook Papers" (internal documents leaked by Frances Haugen in 2021), we know Facebook lacked resources in the relevant local languages, Tigrinya, Amharic, and Oromo (with about 80 million speakers combined) [198]. Only a year into the deadly conflict did Facebook announce they were adding local language support in the three languages, as well as Somali [193]. At least one instance related to the violence was heard by the Facebook Oversight Board[3] [199], but it did not change the material conditions of Facebook's policy in the region immediately [200].

Facebook's content moderation failure is part of a larger pattern of universalism. The African continent has a population of more than a billion people, about 20% of them active Facebook users[3], but until 2021 had only a single Facebook corporate office (in Johannesburg, South Africa) on the continent (which lacked any engineering teams[4]). The office representing the headquarters of Facebook's African operations was in Ireland, a continent away, until 2015. While HQ location was likely motivated by tax incentives, profit of the cybernetic loop (the Facebook product) was more important than having the loop be properly grounded in local context with more localized input on the development teams. By comparison, the active user population of the United States[5] and of the pan-African community are roughly the same, but there exist 37 Facebook offices in the US. While there are more than 50 nations on the African continent, each with their own complex, rich history, till 2021, there was only one office on the continent. Here we see the profound failure of the libertarian cybernetic product. A product developed and attuned to one context (the US) fails to generalize globally, with deadly effects.

Where was the humility? DS and AI are designed, built, and deployed by a small group of people with limited lived experience and conception of potential implications and differential effects. Diverse lived experience alone is not sufficient to fix the problem; instead, it is the embrace of the diversity beyond one's own experience, and the experiences beyond those on the development team. Here is where we find Facebook's methodology and discontents: for the majority of Facebook's operations on the African continent, there was not a single engineering team located on that same continent. Now, we can see that the cybernetic loop of the DS profession is lacking that all critical mechanism: feedback-as-input. Whatever the precise conceptual process that takes place for individual data scientists, the implications of the practice are clear. Not only is there a lack of humility (facilitated by feedback-as-input)—as called for by Miceli [201] — but bias becomes inherent features of systems where it has no place. Further, data becomes subject to what Benjamin calls

the "allure of objectivity," [202] appearing as an apolitical, uncorrupted artifact that yields unquestionable results .

While Facebook represents an extreme application of universality to a combined DS & AI system through its location and language politics, it is not unique. There is a kind of universal design at play, albeit the conception of users is unclear, but clearly centered on a small portion of the global society. Proponents of information and data exchange have long peddled universalism as a core value. Chan discusses how "the colonial underpinnings of global knowledge production continue[d] to reassert Western frames of thought as universal scientific truths" [203] in cases of computing and specifically information exchange . Calls to both diversify the way we understand critical data studies and AI ethics, as well as integrate non-Westernized ways of knowing explicitly target "universalism" as responsible for our current, Western-centric "interpretations of datafication" [204]. Coleman argues the assertion of universalism in data is colonialism in a new form, or *digital colonialism* [205]. As universalism appears baked into the bedrock of information exchange theory, identifying the point at which it became the predominant mindset informs how we might go about eliminating it. What humility can help us fill in is the *social valences of data*, or the ways that data meanings might be mediated across varied social settings and applications [206].

3.1.3 Pinpointing the missing tenet of DS & AI epistemology

How can we make sense of the ethical obligations of DS and AI practitioners? These professions play an important role in organizational oversight and perceived effectiveness as they study system-user interaction; we know that data system problems rarely hold self-evident solutions and require a great deal of discretion on the part of the practitioner(s) addressing them [207] and that the continuation of any system is subject to ongoing sociotechnical negotiation [208]. **This is not to say that DS & AI practitioners operate without oversight, but rather that their organizational role gives them a sense of sys-**

tem oversight, while still subjecting them to existing corporate hierarchy and normative organizational expectations.

This dual role may explain, in part, the data-practicioner-as-whistle-blower phenomenon that has appeared multiple times recently as has Cambridge Analytica whistle blower Christopher Wylie [209]. The overlap between the data professions and whistle blower population is likely not coincidental; professions in the technology—or technical infrastructure—rich cultures of post-industrial nations are now being confronted with the implications of the systems they help design, build, and deploy around the world. A whistle blower, in a digitized society, engages in period of humility [210]: they must evaluate their work, the goals of the organization they represent, and the implications—particularly, the harmful ones—on a greater population or community.

This kind of humility is rare in a discipline defined by both universal *applicability* and *utility*, and the lack of humility—as reflection on self by the practitioner—is a clear cause for concern [201]. The importance of ethical humility has been well studied, particularly as it pertains to research ethics [211], but it belongs too in applied DS practice. Agre's critique of the lack of humility—as reflexivity—in the field of artificial intelligence [212] was built on Suchman & Trigg's description of the "overlooked" reflexive relationship between the designer and user of a system [213] and Woolgar's claim that all systems are embedded in a historical and institutional context [214]. When reflexivity is missing from DS and AI workflows, solutions and systems are built with a lack of awareness for the contexts in which they are to be deployed. Further, there is a lack of awareness about how well the practitioners designing those systems understand those contexts and their complexities.

There are a host of toolkits (both for design and implementation), practitioners' guides, and checklists that promise to help data scientists and AI developers recon with the societal implications of their products. However, going back to the assertions of cybernetics – that fundamentally, all systems are quantifiable—and the deeply libertarian strain of the field that emerged in the US, these aids are stopgaps for a much larger epistemological chal-

lenge. In 2002, Agre was already pointing out that the field of AI had "no effective way of recognizing the systemic difficulties that had arisen as the unfinished business of history has interrupted in the middle on an intendedly ahistorical practice." [212, p. 131]. Our mission in this work is to bring back a sense of history – specifically, historical precedent in cybernetics—to DS and AI. While the design aids mentioned above *do* help practitioners reason about choices they make in system implementation and execution, we argue that going beyond reflexivity into upending the contemporary practice of DS and AI at an epistemological level will require *humility*. The tradition of reflexivity in computational practice is practiced on a personal level, with individual practitioners explaining their relationship to their subject(s) [215, 216, 217]. Humility, we argue, goes farther, suggesting that the premises of the fields in question—DS and AI—are fundamentally faulty. If we want to continue practicing them in the predominant strain (that of major tech companies), we will have to both re-envision and revise the epistemology of these fields, which begins with critiquing their precedent, the US strain of cybernetics. The theory and aesthetics of this US strain have played a fundamental role in the development of what we now understand as the cultural issues of the Silicon Valley tech sector.

3.1.4 Tracing C^2 (DS) and C^3 (AI) back to C^1

To be clear, cybernetics is not wholly at fault for the failings of DS and AI systems. Ironically, cybernetics is conceptualized as a regulatory system, one that is based on biological systems, and can be adapted to any number of applications. Eden Medina, for example, describes how cybernetics was used in the Allende's Cybersyn program, as part of the socialist regulation of Chilean enterprises [218]. Slava Gerovitch details how cybernetics became a language of "objectivity and truth," upending the former *Newspeak* (Soviet ideological language) and contributing to a post-Stalinist era of Soviet science [219].

The term *cybernetics* resembles *cybernétique*, French physicist André-Marie Ampère's sciences of government in his classification of human knowledge [220]. Ironically, US

cybernetics is distinctly anti-regulation & anti-government; cybernetics is a study of regulation, where desired output is a stable system in which one of a handful of desired states are achieved (think healthy forest or thriving *The Sims* population [221]). Decreasing variety in potential output states (achieving one of finite desirable outcomes), necessitates the system being formulated with a varied knowledge of possible scenarios. This is the *law of requisite variety*, in which only varied system input and knowledge can eliminate variety in potential output states [165]. When variety in setup, knowledge, and exposure are eliminated, problems arise.

In short, cybernetics done right requires all incantations of variety. While cybernetics is not tied to any ideology, however, the version that emerged in the United States was coupled with a staunch technolibertarian emphasis. This outlook rubbed off on a number of (in)famous figures in the modern computing movement. Furthermore, it is this strain that shaped contemporary DS & AI systems.

What cybernetics achieves Cybernetics is widely attributed to Norbert Wiener, among others. Fundamentally a way of enacting a regulatory system, in which continual feedback is reacted to by the system with a series of micro-adjustments, cybernetics relies on the principle of *order from noise* [222]. As a regulatory model that can be applied to many fields, ranging from ecological to technical, cybernetics mimics biological systems (e.g., the human central nervous system). Another key concept in cybernetics is *autopoiesis*, referencing the self-maintenance chemistry of cells, through which a system can both produce and maintain itself, usually by generating its own parts[223].

DS is thus second-order (C^2) cybernetics, where the human data scientist(s) serve as human observers in cybernetic system model construction[165]. AI (especially generative AI) is one step beyond, constituting third-order (C^3) cybernetics. Machine learning (ML) has long been acknowledged as having cybernetic origins [165], while AI takes after the general *homomorphism* (identical mapping of representation in one model to another [224])

of cybernetics, where objective world structures are not the basis of judgment. Instead, the basis is the type and order of phenomena, as understood by the system [165]. ChatGPT, for example, has been classified as *bullshit* (in Harry Frankfurt's terminology [225]), given that it produces not semantic knowledge, but merely employs complex statistical frequency counts to determine the most likely next word in the sequence [226]. ChatGPT completely neglects ideological nuance or semantic similarity; instead, it functions on statistical frequency, handily pointing us back to AI as a type of C^3. Sherry Turkle also describes AI practice as having begun with the goal of "finding out how to program machines to do certain acts people do," [227, p. 234] which is perhaps the ultimate cybernetic goal.

Many DS & AI problems are essentially those of *complex adaptive systems* [228] and cybernetics is a way of "marking out what is achievable (for probably many of the investigations of the past attempted the impossible), ... then providing generalized strategies, of demonstrable value, ... [to] be used uniformly in a variety of special cases" [229]. The core of cybernetics is the *cybernetic loop* where system input is treated as impetus for action. Every feedback-as-input consequently produces new feedback, creating a ouroboros effect.

Here we find a missing tenant of DS & AI's epistemic cultures. Generalization suggest that while no system can respond to all situations, any system with significant exposure (via relevant data) can address the relevant problem(s) to which it's assigned. This parallels to the notion of *abstraction* in computing—or that a system, once built, can generalizable to similar scenarios [230]. Abstraction becomes *universality*: in abstracting the nuances of a data point, a new politics is imposed. In this universalist politic, all data points represent uniform images of events and individuals. Geoff Bowker described the US strain of cybernetics as creating a new language with the appearance of "legitimacy exchange", which encouraged a new kind of universalism [231, p. 116].

However, we know cybernetics (as a study of data systems) as universal is insufficient. There is a close relationship between design and data, as Feinberg describes [232], making concerns about universalism in design parallel those in DS & AI. Hsy argues that

universality in design can result in overlooking individual situations [233], while Ryhl illustrates how attempts to have designers empathize with disability results in tropes and increased stigma regarding the target community [234]. Hamraie posits that assumptions about the ideal ("good") version of a design that attempts to remove or disappear, rather than work with disabilities [235]. Human-computer interaction (HCI) struggles with universalist processes and methods where "assuming the ability to design for one user" leads to the "exclusion of many others" [236].

3.1.5 The libertarian, Northern California strain of cybernetics

Cybernetics is inseparable from the image of the *computer bum.* Ensmenger summarizes Joseph Weizenbaum's canonical critique [237] of the *bum* as a young man who embraces computing as an end, rather than a means to an end [238, p. 39]. He (never *she*) is "myopic" and "socially maladjusted" [238], never taking his eyes off the screen. The *bum* is the embodiment of the libertan approach to cybernetics. What is beyond the screen appears knowable in conceptual familiarity, while being simultaneously actually unknowable—a world the data scientist has not encountered. Indeed, Stewart Brand, a leading intellectual figure of the personal computer era, riffed off Weizenbaum's derision for the computer bum for the title of his article celebrating the advent of the modern computer as "good news, maybe the best since psychedelics." [239] Brand subsequently described computers as capable of "suppress[ing] our animal presence" and that when you communicate through a computer, "you communicate like an angel." [240].

Brand is also perhaps the figure most associated with the New Communalist (NC) movement in the 1960s-80s, a movement that inspired a number of prominent figures in the personal computing revolution. Unlike many of their peer counter-cultural movements, NC followers eschewed "agonistic politics" [241] and focused on building new model communities. These communities should be characterized by the transparency and accountability the NC community felt missing from regarding Vietnam and Cold Wars (though Brand

himself had considered a career in the military [242]). The mountains of Northern California were chosen as a place to launch the anti-hierarchical and anti-bureaucratic society. Here, the political focus was on "cybernetics and systems theory" that disavowed "vertical hierarchies and top-down flows of power" [243, p. 38].

Brand's *Whole Earth Catalog* (WEC) was the movement's bible, catering to the NC. WEC catered to Brand's distrust of centralized authority, selling tools to support both the "leaderless Internet" [244, p. 12] and larger personal computer revolution. This obsession with "rugged individualism" [238, p. 50] ironically was based not on the individual liberties of all, but in the omnipotence of the individual NC member. The ethos of the WEC was clear from Brand's editor's note in the first issue:

"We <u>are</u> as gods and might as well get good at it. So far, remotely done power and glory—as via government, big business, formal education, church—has [sic] succeeded to the point where gross defects obscure actual gains. In response to this dilemma and to these gains a realm of intimate, personal power is developing—power of the individual to conduct his own education, find his own inspiration, shape his own environment, and share his adventure with whoever is interested. Tools that aid this process are sought and promoted by the WHOLE EARTH CATALOG." [245].

Critically, WEC espoused connectivity between individuals, as demonstrated by Brand's mention of "shar[ing]" a person's adventure with "whoever is interested" [243, p. 82]. The emphasis on mutual interest is key; a predominant feature of the community that gathered around Brand and his publication was a belief that the small, elite group to which they belonged had attained a level of independence and self-sustainability that could be—at a later point—adopted by the larger population. There was never an emphasis on outreach. Rather, ideas were shared within an existing cadre of like-minded individuals. Later outshoots of the WEC community, including the WELL (Whole Earth 'Lectronic Link) internet forum and Global Business Network (GBN) consulting firm strengthened bonds between this community [241].

Understanding the universalist undercurrent of contemporary data systems practice requires revisiting the early days of Silicon Valley and its passionate embrace of cybernetics. As a field of study, cybernetics had a strong imprint on a generation of pivotal characters in the story of modern computing, many of them inspired by Brand's espousal either directly or through WEC. Brand's own enthusiasm for cybernetics as a way to "reject hierarchy in preference to the idea of a networked world in which the center of control constantly moves to wherever is most 'relevant and useful' " [240]. The *libertarian* outlook arose from Brand's belief in the computer as "a true opportunity for individual empowerment" as opposed to the previous perception of it being "a tool of bureaucratic control"[246]. However, the cybernetic loop represented by contemporary DS is broken, and some of that breakage can be traced to the liberalist influence on cybernetic practice.

The libertarian ethos so fundamental to modern computing's identity [172] (including the "Californian Ideology" [247]) was fundamental to the spread of cybernetic theory. Borsook characterizes the Silicon Valley understanding of libertarianism as both a world view and political philosophy, which includes "all different colors of free-market/antiregulation/social-Darwinist/aphilanthropic/guerrilla/neo-pseudo-biological/atonmistic threads" [172, p. 8]. The resultant *technolibertarianism* combines the classical anti-regulatory & pro-market bias with a philosophical aversion to "human connection" and "a discomfort with the core of what many of us consider it means to be human... It's an inability to reconcile the demands of being individual with the demands of participating in society." [172, p. 15]. In short, technolibertarianism is a way to understand the world only by virtue of self; it is naturally congruent with universalism, enacted from the practitioner's own standpoint.

The technolibertarian approach to technology is one of equalizing differences, premised on an idea that technology erased differences between people, premised on these universalistic understandings. Cultural differences become meaningless in the face of technology: technology meant to bridge differences instead erases them. In 1995, Nicholas Negroponte (MIT Media Lab founder) rhapsodized about the Internet's ability to "flatten organizations,

globalize society, decentralize control, and help harmonize people" [243, p. 1]. In another exemplar of techno-optimism, Negroponte famously stated "We'll take tablets [from the One Laptop Per Child project] and drop them out of helicopters ...then go back a year later and see if the kids can read." [248, 249]. Eschewing the local context in which the tablets would be deployed, Negroponte still claimed the tablets would change lives. Ironically, the early Internet was in fact a product of those institutions Negroponte likely sought to flatten. ARPANET was funded by the US Department of Defense [250], and the US computing industry was shaped by early and robust investment by the US government, which greatly expedited the development of computing technologies [251].

Internet policy was not immune, either. The landmark 1996 *Telecommunications Act* was architected by outspoken libertarian and then-Speaker of the House, Newt Gingrich. The Act offered novel deregulation to telecommunications firms and included the now hotly-contested Section 230 of the *Communications Decency Act.* This act of deregulation made telecommunication companies, especially Internet service providers (ISP), effectively immune to civil liability charges stemming from actions users (third parties) took on or through their services. This laissez-faire ("let be") approach mirrored the argument of John Perry Barlow's *A Declaration of the Independence of Cyberspace*, in which Barlow decries the "tyrannies" the US government seeks to impose on cyberspace—ironically, the very development of which it funded [252]. In the Declaration, the Internet is described as a self-regulating organism that should exist outside of state control. The Declaration has its fair number of critiques, including Lessing's [253] position that cyberspace was moving towards "an architecture of control" and Bomse's positioning of Barlow's argument as merely a new manifestation of an existing tradition of historical American liberalism [254].

The position expressed in the *Declaration* is consistent with Barlow's rage against what he understood as the US attempting to establish "imperial control over cyberspace" [255]. Raging against the Clipper Chip, the US's failed attempt to establish a de-facto backdoor to consumer telecommunications devices, he claims having never before been scared of

government telecommunications control ("It's always seemed to me that most governments are too incompetent to keep a good plot strung together all the way from coffee break to quitting time") but that "nerds like us" (readers of Wired magazine) must spread the message to the unconverted (those "who have never touched a computer") and for whom "it is a revolutionary war we are engaged in" [255].

Universalism as early Computational Universalism The NC community was both homogeneous and powerful. Brand himself inspired and associated with a number of so-called computing icons. Bob Albrecht, of the Portola Institute, and WEC patron, later became a co-founder of the Homebrew Computer Club (HCC) [243, p. 69–71]. The HCC launched the careers of a number of luminaries of modern computing, including Steve Jobs. At his 2005 Stanford University commencement address, Jobs reflected on the WEC: "When I was young, there was an amazing publication called *The Whole Earth Catalog*, which was one of the bibles of my generation ... like Google in paperback form, 35 years before Google."[8]

Bill Gates published his infamous "Open Letter to Hobbyists" in the HCC newsletter, in which he expressed frustration with hackers' copyright violation behavior, including against Microsoft (then Micro-Soft) products. The argument became a landmark for protecting the intellectual property (source doe) of tech companies. Members of the HCC developed the "hacker ethic" made infamous by Stephen Levy's 1984 study ("Hackers: Heroes of the Computer Revolution").

The WEC network included Douglas Engelbart and Alan Kay, widely regarded as the "fathers" of HCI. Engelbart kept an office around the corner from WEC and Kay described WEC's do-it-yourself ethos as a way to deal with "complicated ideas" through computer models [243, p. 112]. The larger WEC community never branched out far from the homogenous NC movement that beget it. Despite allegedly divorcing themselves from the society they despised, NC members still managed to transfer their social biases and stereo-

types. Membership consisted primarily of the young, able-bodied, White sons of the professional class masquerading in the garb of the working class.

Childreading and housekeeping once again became women's work. Women's liberation received minimal coverage; Brand selected only a passage about masturbation from his friend Diana Shugart's *Our Bodies, Ourselves* and cautioned his (presumably) male readers not to view it as "a marriage manual" [243, p. 98]. The American Indian Movement and Black Panther movement made little mention; the singular feature in the WEC's history that explicitly discussed race was a 1970 "Black Panther Reading List" [243, p. 97]. WEC was intended for those who were "masculine, entrepreneurial, well-educated, and white" [243, p. 97]. In particular, the WEC and NC environments were distinctively masculine, which matches the universalist critiques that Braidotti describes [40].

WEC critiques mirror those of contemporary computing culture. WEC espoused universalism: a do-it-yourself attitude could affect fantastic global change. Connecting all members of society is a noble goal, but the WEC community was not particularly interested in connecting with the "other" [256]. Instead, WEC-ers sought the creation of a new elite within the US, based on perceived aptitude and ability. This specific embrace of cybernetics was both a means of fighting elitism *and* developing a new elite.

3.1.6 Approaching an epistemological re-imagining of AI and DS with humility

How might we rebuild cybernetic practice, implementing that humility Borsook dreams of? Consider Brand's 2007 amendment to his infamous 1968 statement ("we are as gods and might as well get good at it") [245, p. 3]: "We are as Gods and HAVE to get good at it." [257, p. 2]. The fields of DS and AI are still struggling with "getting good at it" almost fifteen years later. Despite the computational micro-improvements that delight some AI & DS practitioners, the list of AI and DS scandals increases daily, with fundamental tenets of human health and happiness tirelessly encroached upon. We are constantly reminded of lacking professional humility, every time a DS or AI product fails to "generalize" (betrays

its universalist design). These systems should support economic and social goals, allegedly contributing to overall human health and happiness. Instead, they are some of the most destructive *weapons* against these goals, using Cathy O'Neil's moniker, "weapons of math destruction" [258].

Compare Brand's revised statement to what Haraway famously argued: "I would rather be a cyborg than a goddess" [168, p. 116] . The cyborg (and the cybernetics-critical strain of cyberfeminism) can serve as an alternative to our current paradigm of cybernetics. Recognizing that our AI and DS systems are fundamentally entangled with society (adopting the cyborg stance), allows us to potentially find ourselves in a wholly different computational environment. As J.A. English-Lueck, anthropologist of Silicon Valley's technorati reminds us, "Silicon Valley has its own complex pivotal technologies, devices that structure the lives of people, acting as critical capital in production, shaping relationships between people, and providing a symbolic framework for their understanding of the world" [259, p. 53]. The challenge is neither small, nor holds trivial implications. However, cyberfeminism and cybernetics are not ideologically distant; Sadie Plant, early proponent of cyberfeminism, was head of Warwick's Cybernetic Culture Research Unit. Cyberfeminism as cybernetic critique is also an echo of the interest and desire to build interesting systems, albeit systems that are attuned to the site of application.

A cyberfeminist approach to DS and AI systems fulfills Braidotti's call to turn to creative avenues, such as cyberpunk, "to find nonnostalgic solutions to the contradictions of our times" [40, p. 349]. Following Braidotti's advice, we propose an investigation of cybernetics, one that allows us to see how the theoretical infrastructure of many cybernetic-incantation systems we encounter today, had their universalist tendencies exacerbated by masculine libertarian sympathies. In turning a nonnostalgic eye to cybernetics' legacies, we employ history in pursuit of a better future.

Barlow & Brand practiced a version of computing that erased its histories (e.g., the US government's role), while Negroponte espoused connectivity at the cost of culture (e.g., dis-

tributing computers requiring English commands). Where WEC struggled to imagine the "other" (non-man, non-White, non-wealthy), cyberfeminism offers a standpoint based on the cyborg. Linda Dement offers this standpoint as "a new politics of feminism", per Kay Schaffer [260, p. 162]. The cyborg politic can embrace the expansiveness of experience and culture [31] which is sorely lacking from current AI & DS practice.

Ultimately, the professional communities of DS & AI must come to terms with the fact that their epistemology is joint with bias—as nuance reduction or obfuscation—that riddle DS & AI systems. Advocating *technosymbiosis* as a more-than-human ontology of the relationship between AI and its environments, N. Katherine Hayles asks, "how did the field of data communication arrive at conventions that prioritize rationality and devalue emotion?" [261, p. 13]. By prioritizing universalism, as the system design methodology beget by universalist understandings, DS & AI ignore the nuance requisite to system deployment in a given context. To address these nuances—namely the potential discriminatory biases that might possibly result—a new epistemological practice should focus on system development where systems are built context-aware and situationality-respective. Designing and deploying in this way requires a fundamental shift in how we conceive of DS & AI as professions and fields. If we continue with the exclusively-cybernetic paradigm, we must reckon with the legacies of US liberalism and its inherent implications.

Returning to the cybernetic law of *requisite variance*, an alternative epistemology could include re-thinking who belongs in the AI & DS communities [262]. Rather than seeing AI and DS infrastructure as universal purpose tools, cyberfeminist epistemology of these fields attune us to the use of these tools in context. What if we built DS & AI system development teams that deprioritized the "god-like" AI or DS professional, favoring instead local community or domain-specific experts, as practitioners of contextual understanding and care. How might these *cyborg* teams respond to the challenges before them?

3.2 Uncovering Data Workers Beyond Data Scientists & AI Developers

In the prior section, I discussed how data science developed as a practice. Namely, it was, and continues to be, a practice with generalizability baked into its core; in data science, universalism can be traced to the early practice of cybernetics, and today it continues through the belief that generalizable systems, which abstract over local context and culture, are possible. Given that the practice of universalism has thus far had significant problems, particularly in the form of AI and ML system deployment in sites where the systems are not fit for purpose, how do other data workers handle issues of contextualization?

We know that participants in data work are not limited to data scientists—like any professional community, there is a diverse bevy of practitioners; there are data wranglers [263] or domain experts who engage in data work to facilitate advances in their subject area [264], alongside data scientists [265, 266]. There have also been extensive explorations of how individuals in other settings, such as nonprofits [267] and healthcare [268] engage with data; however, these works describe how data work—as a secondary task—affects mission-drive work and professionals within an organization rather than understanding the needs and experiences of individual practitioners of data work. We propose adopting the community of practice framework to identify and think about the various groups engaging in data work [269], as the framework yields a way to understand the intricacies of a given community of practitioners. Community of practice dictates that rather than using a central and exterior binary to think of data work, we see membership in the community as a gradient. Groups with various titles and task, like data scientists and data analysts can be considered full (i.e. experienced practitioners) members, and groups like data wranglers or domain experts working with data as peripheral members. Novice participants who, upon gaining more skill and experience, may or may not join one of the groups of full practitioners. Given the variety of titles used for computationally-intensive data practitioners, we isolate only these two terms, data scientist and data analyst. Our goal is to suggest one role that has,

theoretically, experience in working on datasets with diverse subjects (data scientists). In contrast, data analysts focus on a subset of data specific to a given organization or subject area but their work is still computationally intensive. When examining data work from this perspective, there is a disproportionate amount of scholarship related to full participants in comparison to peripheral members.

Why does our relative lack of knowledge about peripheral data workers matter? In the last few years, the CSCW community has come to understand data work as a human-centered process subject to discretion and interpretation [270, 74, 208, 271, 272]. There are even new subsets of data science based on the need for nuance when working with data about individuals, such as "human-centered data science" [75]. Moving the focus from the textual data exclusively to the way that human data workers interact with and process that data has key implications for fairness and transparency in data-driven systems; as a research community, we know that data is never "raw" (and to describe it as such "is both an oxymoron and a bad idea", in the words of Bowker [273]) and must be understood in context [274]. Engaging in active efforts to both document and preserve the context of datasets is an important first step towards designing equitable and just data-intensive systems; however contextualization often falls to the wayside in favor of other facts such as "universality", as Klaus Scheuerman et al. describe in their study of computer vision datasets [275]. Contextualization contributes to what Monroe-White describes as *emancipatory data science*, or "data work that frees members of marginalized communities from being the 'object' to the 'subject' of data science framings" [276]. Monroe-White describes that these practices will help eventual dataset users understand the "decisions regarding why, how, what, when, and where data are collected, managed, analyzed, interpreted and communicated" and that contextualization should involve those who are the subject of the data being used [276]. Loukisass, too, questions the role of "locality" in data [85] in this manner, while D'Iganzio and Klein suggest a feminist-informed handling of data [47].

Even the most minor attributes of a dataset's "texture", or the relationship between

the dataset's infrastructure and environment, has weight on the correct (or well-informed) contextualization [277]. Only by understanding the motivations and actions of the human actors can we begin to reason about the functionality of such systems in their entirety. As there is an increasing call for XAI (explainable artificial intelligence) which requires uncovering the social context of these "human-AI assemblages" [278], the need for tools and process evaluations to support such work is only growing.

To kick-start this investigation of human-centered data work, Muller et al. established several open, key questions about the role of humans in data science [74]. Figure 3.1 shows the original text of these questions, with our summary category tags; throughout this paper we address these questions by their shorthand tags. Answering these questions pushes us both towards a scholarly understanding of these work practices and the development of tools and processes to support a broader array of data work. Our research attends to those who perform data work, but are not data scientists in the common use of that term, or other fairly-full practitioners (e.g., data analyst, data engineer, etc.). These workers are experts in other domains, but they collect, clean, analyze, and share data as an essential part of their work and identify data as a principle task of their role. In other words, they engage in activities that are familiar to data science but they are not, generally speaking, data scientists. We believe attending to these workers and practices is important for fostering a diverse approach to what counts as data work, which in turn aligns with commitments to equitable and just principles of labor and broadening participation in computing.

In this section we[2] describe the results of interviews with 19 individuals engaged in civic data work. We discuss both the specific needs and assets of this group of data workers and why their experiences and methodologies can both inform the tools we build for other data worker populations (such as data scientists) and call for professional tools and systems that address their needs. All work in the civic sector—either directly for local and state government or for non-profits in large cities in the United States. They work on "civic

[2]This section describes work collaboratively undertaken with Amanda Meng, Carl DiSalvo, Britney Johnson, Ben Rydal Shapiro, and Betsy DiSalvo, which appeared at CSCW in 2022.

Table 3.1: Guiding questions for studies of data workers, from Muller et al. (2019):

Question number	Question, as posed by Muller et al.	Question tag
1	How do data science workers approach their tasks? What are their strategies? What can we learn about their views of data and process?	How?
2	What (and who) is missed in our prevailing accounts and assumptions around data and data work? If ”data is never raw”, who does the cooking, and how is this work performed, recognized, and accounted for?	Who?
3	Data science tools are generally designed for one user at-a-time. Other complex tasks have benefited greatly from collaboration practices and collaboration technologies. Are there opportunities to bring some of these lessons to data science? What are the distinctive collaboration needs and constraints in data science?	Tools for collaboration
4	What methods are valid and tractable to assess the usefulness and usability of tools for data scientists in service of iterative design? How can learnability be assessed when domain experts, not programmers, need to learn new languages and tools simultaneously, the learning of which takes far longer than typical usability studies? How can efficiency for skilled users be assessed when the uses of the tools are so diverse that benchmark tasks have little face validity?	User-based tool assessment
5	Who are the consumers of data science work and what are their needs? How can data science processes and tools address these needs, e.g., the level of transparency and comprehension consumers desire?	Implications of data work

data" which Sinders describes as data that reflects "people who live in the communities and cities where the data was gathered" [279] and we extended to include datasets that describe the infrastructure and municipal services associated with those cities. Most of are participants first and foremost domain experts in another field, for example, the main cause of the non-profit they work for, and perform data work that amplifies and extends that mission. Within their organization they would be described as 'the data person', or the individual with the most computational data savvy; they are the primary data workers within their respective organizations, showing the outsize impact of their work on their data ecosystems. When civic data ecosystems function correctly, especially when they increase transparency through open-access initiatives, they have myriad beneficial social implications [280, 281]. The contrapositive also holds: as Irani and Marx describe, "[t]he withholding of public information obscures and obstructs the democratic process by denying ordinary people the right to know what's being done in their name." [282]. Further, given the nature of their work, the clients of these organizations and governmental entities do not have much of a choice when engaging with these services, making them important sites of transparency investigations. Citizens' private data, too, is increasingly becoming public data in the era of revolutionary civic tech—as described by Boehner & DiSalvo [283]—which further compounds the need to understand how these data workers go about their work. Understanding this group of data workers, too, lends us insight about what data work in low and limited resource environments looks like, in contrast to the large technical organizations many participants in past literature were employed by.

Our results inform several of Muller et al.'s questions, with respect to civic data workers specifically, but extend to larger group of data workers more generally. First, our participants shed light on how data work is done in the civic contexts, which compared with high-tech data science efforts are relatively low-resourced (Question 1). We also discuss their applied and function-oriented perspectives on data. Second, we highlight these workers as "peripheral data workers" and demonstrate their role in major data ecosystems

(Question 2). We use the term ”peripheral” in a technical sense to call attention to how the tools and processes of these workers are at the margins of what is commonly considered as the data science community of practice. While peripheral data workers have been missing from much of the scholarship, we argue that they play as consequential a role as their full professional peers, namely data scientists at large technical organizations. Third, we describe their collaboration practices that are uniquely shaped by their comparatively low-resource work environments, replete with outdated tools and little ability to acquire new ones (Question 3). Fourth, we describe the tool acquisition practices of this group of domain experts, rather than formally trained technical workers; namely, that given their work environments and educational background, they are limited in choice of, and training for, new tools (Question 4). Fifth, and perhaps most importantly, we describe the ways in which this group of peripheral data workers goes about contextualizing the datasets they work with and the novel formal and ad-hoc systems they develop to do so (Question 5). These practices are informed by the proximity the data workers have had built into their workflows or have actively sought and or developed with the data.

We close with two insights for the CSCW community. First, a need to continue to expand our definition of who is a data worker and, subsequently, how the needs of data workers might vary between subgroups. We identify and position various subgroups of data workers via the community of practice mapping, which highlights a more complex relationship between full and peripheral practitioners than professional vs aspiring. We further suggest that rather than fetishizing the advanced computational skills of data scientists as model community members (and subsequently positioning them as the community members to emulate), we consider at the ways in which data scientists can learn from other members of the data work community, particularly with regards to data contextualization. Second, by examining one of the ”peripheral” groups of data workers, we describe specific limitations hindering the use of tools and systems by data workers in low and limited resource environments and propose ways to address them.

3.2.1 Related work

The CSCW research community has paid increasing attention to how data scientists go about their work in recent years. Our work furthers that effort, introducing the domain-experts-as-data-workers and their contextualization practices, which hold promise for other, more full participants in data work. Accordingly, our work builds off three areas of work: understanding the role of humans in data work, identifying who participates in data work, and finally, how the currently studied data workers go about their work.

Human-centered data science

While data science—and data work more broadly—has traditionally been seen as a "rational 'data-driven' process of 'discovery' that reveals the underlying nature of a domain" [74], this is no longer the case. Increasingly, we understand the way that humans performing data work shape the understanding and dissemination of the data they work with. Feinberg describes individual data workers as "designers" of data, rather than objective appropriators, highlighting the role the that human perception, background, and motivation play in the way dataset contents come to be understood [232]. Even the act of reading a database (making sense of how information is organized within it) is an act of "awareness, reflection, and control" [284]. Data work is also a collaborative practice subject to social interaction and dynamics: Koesten et al. explore how individuals use a series of patterns of activities in order to engage in data sense-making, describing the cognitive and verbal practices that take place [285]. Miceli et al. expand on the power dynamics that take place in data work, highlighting the case of data annotation in computer visualization datasets; they call it a "sense-making practice" that is frequently influenced by the labeler's higher-ups, whose decisions sometimes conflict with the labeler's factual or logical inclinations [15].

However, much more attention is paid to improving models than improving datasets, even though data provides an upper bound for quality in machine learning and algorithmic systems [286]. One of the biggest issues with data is the lack of proper contextualization.

The implications go beyond reduced accuracy and poor model fitting: a lack of contextualization leaves room for issues of bias and fairness. When issues arise in datasets they tend to result in data cascades, or compounding, negative events that compromise the quality of the data as a whole, as termed by Sambasivan et al. [13]. D'Ignazio & Klein underscore the importance of contextualization practices, introducing reflexivity as a necessity for restoring context in data work. They identify "social, cultural, historical, institutional, ... or material," and who participated in the creation and curation of that dataset as starting points [47]. Loukissas, too, argues that contextualization of data requires an intimate understanding of the environment from which it originates [85]. When the associated origins and context of a dataset become separated from the textual data, it is easier to misuse or misapply a dataset. For example, consider a model to predict increases in home value in urban areas that fails to account for decades of housing discrimination and inequality in parts of those cities.

Miceli et al. identify how this context restoration praxis might take place specifically for large computer vision datasets and the individuals involved in labeling them [201]. There are other contextualization practices proposed for large-scale datasets: Bender & Friedman proposed data statements, essentially a short biography, to accompany NLP datasets [287] and Gebru et al. propose datasheets for datasets which include, among other information, "motivation, composition, collection process, recommended uses" to help facilitate conversation between dataset creators and consumers in the style of fact sheets that accompany electronic equipment [288].

But who does the (data) work?

One limitation to the innovative contextualization practices described above is who they target. While large-scale machine learning datasets are undoubtedly important places for contextualization practices to take place, they are not the only place. Further, these practices are aimed at researchers and practitioners in large technical organizations, who have

both access to a community of skilled peers and ample opportunities for continued training. If we look at the data work as a *community of practice* including, but not limited to, individuals who meet this criteria—such as data scientists who work with large scale datasets in high-tech organizations—it becomes clear there are other groups who could benefit from these contextualization initiatives. They demonstrate new modes of data contextualization and help us refine the contextualization tools and practices we develop for data scientists.

Lave & Wenger define a community of practice as the participants in a shared activity, at varying levels of experience and skill [269]. The community is defined not by the kind of work being practiced—many kinds of knowledge-building communities can be understood as such, formal or informal—but by the "structure and character of [the] community [that] emerges" [289]. Critically, a community of practice has full participants (experienced, highly knowledgeable members of the community) and peripheral ones (those who are learning the craft). Lave & Wenger give examples of communities of practice based on physical craft, e.g. tailors and midwives, who have more and less formal apprenticeship practices respectively. The community of practice construct has been applied in the computing space previously, for example to understand Wikipedia editors [289] or to grow the global HCI practice [290].

When we examine data work as a community of practice, we can make sense of data scientists—and similar data roles that prioritize computational skills over work in the subject domain—as full practitioners. It is these full practitioners who have been the subject of much of the CSCW community's focus, rather than members who may be on the periphery of data science but are full member of a more inclusive data work community of practice. Data science pulls from several closely-aligned fields, including information science, statistics, computing, and knowledge domains [266] and peripheral data workers are likely to reside at the intersection between one of these aligned fields and data science proper. Peripheral data workers engage in some subset of the full member's practice, while constantly becoming closer to full members as they hone the skills prized by the data work commu-

nity. Data wranglers are an example of another group of the periphery of data work, as they focus on the early cleaning and organizational steps of data work [63, 263], while full practitioners—e.g., data scientists—have experience with the entire data lifecycle. Previous research has shown how individuals outside the data work community could join it. For example, young children have become data science apprentices, learning the skills of data production, analysis, and consumption [291, 292, 293, 294]. We do note that these are not professional practitioners, rather they are hobbyists or early-stage learners. Content moderators are another such group: while some do so at a professional level and have little agency over what happens to the data after they pass it off [295], others are volunteer moderators, such as [296] have more agency over the platforms they work with, but are usually not receiving compensation or professional evaluation for their work. Feinberg et al.'s reproducibility workshops for scientists from domains besides computing demonstrates a more professional apprenticeship experience [297]. Other professionals encounter data work as part of their profession, though it is not their main role, e.g., health care workers and the challenges they incur while doing so [268]. The data workers we discuss are unlike these previous groups in that they are not necessarily trained in data work, but perform it as a part of their compensated, professional role. Further, they consider data work to be a primary component of their role, but they rely heavily on their subject area expertise.

Building a broader understanding of data work as a community of practice may also address issues of collaboration. The work of Feinberg et al. also highlights the division between domain experts and technically skilled workers; those with extensive domain expertise have spent their careers learning the ins and outs of their subject, while groups like data scientists have spent theirs expanding their technical skill. Bringing the two groups together is not easy. For example, data work produces friction in the roles of health care providers [268]. According to Crisan et al., full practitioners of data work, who may have a variety of job titles, have varying levels of domain expertise with which to perform contextualization work [298]. When external domain experts are brought in to consult on data

work projects, however, the process is challenging and often fraught with technical and time limitations, as described by Mao et al. [264]. Therefore, the experiences and needs of those performing data work on the periphery differ from those of full community practitioners, including data scientists.

Most insights we have from ethnographic or human-centered studies about data workers, which inform this kind of contextualization work, focus on these full practitioners. For example, Pereira et al. study data workers in diverse fields and find that they have similar tool needs and challenges despite different domains, but their participant population all have focused degrees in data work and all but one have titles that explicitly include the word "data" (e.g., data scientist, data analyst, or data architect) [299]. Wang et al. provide insight on the way data workers use computational notebooks, but their survey population is comprised of students in data science or computer science degree programs, or professionals with an extensive technical data science or computer science background [300]. Zhang et al. uncovered the depth of collaboration in data science, but their respondents were a self-selected population of data workers at IBM [301]. Bopp et al. are an exception to this; they explore the impacts of data work on non-profits [267]. However, they find that the focus on data dis-empowers organizations, while our respondents are part of organizations that have long embraced data and do not struggle with its role in their organization.

This oversized focus on highly skilled technical workers as the targets of data work ethnographies and tools is not surprising. Data science and its practitioners have demanded attention in recent years from all corners of society, let alone this research community. However, these highly skilled technical workers are both a limited portion of a much larger population of data workers, and they have been only recently subjects of contextualization inquiry. We theorize that this is the result of data science being used interchangeably with data work in common parlance, as well common definitions of a data scientist's required capabilities listing domain expertise as a final add-on or one skill among many others [265, 302]. This de-prioritization could be due, in part, to variability between domain expertise

required by different roles and difficult defining it. For example, the EDISON framework developed to create a common definition of data science skills and competencies prioritizes technical skills over domain knowledge [303].

We postulate that many of those engaging in data work—perhaps even the majority—would not be considered full practitioners of data science. Data wranglers, for example, would be seen as peripheral at most to the data science practice, however they have a much fuller role in the data work community. All groups that are peripheral to data science are part of the data work community, in which they may still be peripheral or full participants. Peripheral data workers are to full data workers as paramedics are to physicians; paramedics perform applied, acute care for a subset of maladies, where physicians are responsible for a broader array of conditions and conducting further research into the state of the art. In many cases, these groups of peripheral data workers are first and foremost domain experts in their area of expertise and secondarily practice data work as it pertains to their domain. They are more likely to work in low or limited resource environments and, corresponding to their domain expertise, are often in close conversation with those involved in the data's origin, or they are directly involved in that process. This leaves us with two questions, which our work will begin to answer: 1) how well are other groups of data workers, such as those on the periphery of data work, engaging in contextualization practices, and 2) do they have contextualization practices that could be adopted by more full participation in data work (such as data scientists)? Further, given our focus on this group of peripheral data workers, we highlight the tool needs of this specific group and those in similar low or limited resource technical environments through design suggestions. These are pertinent questions especially in the face of the increasing automation of data work, for example, the AutoDS project, and the likelihood that they be widely adopted [304, 305, 306].

3.2.2 Methods

To understand how those engaging in informal data work approach and understand their role, we interviewed 19 individuals between October 2018 and July 2019. These participants work in urban areas in the United States either for local or state government or other civic organizations. The interview protocol was approved by our Institutional Review Board and all participants consented to the study. Our process was based on the grounded theory approach [307].

Researcher Positionality

The authors on this paper are academic researchers who have worked with civic and non-governmental organizational data through assorted research projects and industry positions. We approached this project as researchers on the DataWorks project, a work-training program to broaden participation in middle-skill data work; our intention was to understand how data is collected, organized, and processed in civic and non-governmental organizational contexts—or, environments outside of data-centric or data-focused technical organizations.

Participant selection

The criteria for participant selection included individuals who described their job as data work, or who would refer to themselves as "the data person" within their organization but did not work for a for-profit company. Specifically, we sought individuals those working for local governmental and civically aligned non-for-profit organizations. Potential participants were identified through the research team's local connections and network. Additional participants were identified through snowball sampling [308], with early participants suggesting others. Recruitment happened through email and word of mouth. Participants were not compensated, as many were local or state government employees and could not accept compensation for interactions that took place while they were in their professional

capacity.

Interview structure

During interviews participants were asked a series of questions following a semi-structured interview style [309] and asked to describe their desk set up and asked to draw a diagram of example data flow within their organization, with these last two tasks styled as situated observation [310]. Interviews were recorded and transcribed. Interviews lasted between 28 min and 89 minutes with a total of 14.4 hours of interview recordings analyzed. Four of the interviews were conducted in pairs with two individuals who work on the same team; P8 and P9 were interviewed together, as were P10 and P11.

Data analysis

Transcripts were analyzed pursuant to the open coding framework [311] by the research team through multiple reviews and revisions. First, members of the research team reviewed different transcripts and discussed themes that emerged to develop the initial code book. Second, the first author then reviewed all of the transcriptions, began coding them with the initial code book and returned to discuss with the research team potential modification and nuances to change the code book based upon findings.

Third, after themes were thus refined, one researcher the first author coded all of the transcripts into four themes. Fourth, then two researchers the first and last authors reviewed each excerpt that was identified for each code and discussed how it applied. If there was disagreement as to which codes were applicable researchers discussed and refined codes to help with clarification. Fifth, the first author then applied the final version of the code book to the full corpus, with the last author providing peer review on random excerpts from each code to ensure continuity and reviewed excerpts that were difficult to code. Sixth, the research team then collaboratively identified themes within these the set of excerpts for each code, following thematic analysis [312].

Given the open-ended nature of many of the interview questions, responses went in multiple directions. In recognition of this variance, we generally avoid reporting quantitative metrics about participant responses and instead highlight reoccurring themes within categories of responses.

3.2.3 Findings

We separate our analysis into two sections, namely participants' backgrounds and then observations of how they go about their work. This division roughly corresponds to the dual purpose of this paper. First, we identify this group of dataset domain experts and their ability to contextualize datasets. These contextualization practices can be used to work against bias in automated systems by documenting limitations of the dataset. Second, we observe their work environments and experiences with civic data and provide insight about how this community of tool designers and ecosystem ethnographers might best serve these under-served members of the data science community.

When discussing participants, we purposely choose to obscure their identities as much as possible, given both the sensitive nature of their work and to allow them to be critical of the environments in which they operate. Therefore, individuals' backgrounds, education, and demographics are reported in aggregate and we use the gender-neutral "they" pronoun throughout.

Participants

1. *Participant backgrounds.* Our participants vary in gender and race from national averages in computer and information technology workforce participation [313]. Roughly 30% of our participants are female-identifying, 70% male-identifying, with the former slightly higher than national averages. Participants identify as Black or African American (n=7), Asian American (n=1), and White (n=11); this varies greatly from national averages. They mostly (80%) hold graduate degrees in their

Figure 3.1: Participants' backgrounds, in terms of technical skill—as defined in traditional data science verbiage—and domain expertise. Points 1 and 2 are red to denote that they are comparisons to our participants, where all participants points are dark blue.

domain areas—for example, a Masters in Public Administration for someone who works in government—and the remainder hold Bachelors degrees in their domain area.

In Figure 3.1, we plot the relative domain expertise and technical skill of our participants, in comparison to one another. The plotting methodology is not precise; rather we aim to demonstrate the relative grouping of participants, taking into account their educational background, demonstrated and self-described technical skill, and any previous work experience or history mentioned. Domain expertise relates to formal knowledge of a space gained through explicit educational experience, or self-described history within a specific organization or overarching body (for example, the municipal government). Technical experience is comprised of both formal training, such as a BA in computer science, alongside any skills learned on-the-job, in this specific role or a previous one. In our ranking of technical skills, we rate the technical complexity of tools used, too—for example, performing data cleaning and visualization with GIS tools is considered more skilled than collecting participant responses

through Google Forms surveys. While both types of skills are important and require much practice and experience, we refer to the commonly accepted definitions of technical skill for traditional data scientists as our measuring stick, which prioritizes a particular suite of tools. Therefore, our metric measures technical skill and does not necessarily reflect professional experience and capabilities in a broader sense, so much as they do the metrics by which data scientists are traditionally judged.

As shown by Figure 3.1, our participants tend to be domain experts who have secondarily acquired the technical skills necessary to do data work in their primary domain of expertise. None of the participants are employed by primarily technical organizations. Their employers range from local and state government to small nonprofits and public utilities. Notably, seven of the participants have titles that include "data", "analyst", or "analysis". Only one participant describes their work even indirectly as "data science"; this participant is also the most technically-skilled member of the participant pool. Some work on primarily technical teams; for example, P19 manages a small team (less than 5 other individuals) of GIS practitioners in an organization that conducts regional policy analysis. Others are the primary data person within their organization; P18, for example, moved into their role as the organization's data manager after having gained a reputation within the organization as an "Excel nerd".

2. *Participants' perspectives on data work.*We asked participants a series of questions about their perspective on data work, to capture what similarities and differences they have with the full participant peers, namely data scientists. As participants focus on different parts of the data lifecycle, it is not surprising their responses are varied—from the logistics around each activity (e.g., designing questionnaires and writing guides for focus groups) to analyzing the dataset. The two most frequently mentioned tasks were data collection and cleaning, which were mentioned by five and four participants, respectively. Several participants mentioned the challenges of data collection, as they had to find and integrate data sources from several offices or

departments. P13 describes this process when answering what part of their job takes up most of their time:

> P13: *Collecting. So like pulling data together from different departments. Yeah, I would say collecting.*
>
> Interviewer: *Is that mostly what you mean by that? Just figuring out where I get this data set.*
>
> P13: *Exactly.*
>
> Interviewer: *And asking folks for it?*
>
> P13: *And coordinating with people on to get needs met. So it's kind of like a project manager function. Pulling the people together and getting it in that useful format is probably where the bulk of the time goes.*

For P13, collecting the data encompasses not only obtaining the dataset, but coordinating cleaning that dataset to render it usable. P19 goes into more detail about the all-encompassing nature of cleaning: *"I kind of consider collecting like intrinsic to the job, so I don't think about it being an action. It's just what we do. I don't count that time really. Cleaning is a pretty big deal."* P8 elaborates on the hurdle that messy data presents to project progression:

> Interviewer: *You said you spend most of your time cleaning. Is there something that you would rather spend more time on?*
>
> P8: *I'd like to spend my time more analyzing. But see, ... we don't have great information for a lot of things, so I'm always trying to figure out to the extent of how much information is user providable and how much I may have to ask you go and clean it up... I want to spend more time analyzing.*

Here, cleaning again gets in the way of performing other data tasks the participant would prefer to engage in. P4 sums up their feeling on the labor to reward ratio of data cleaning: *"it's kind of like finding a needle in a haystack."* Cleaning comes up, too, as a task participants find particularly tedious. In the words of P7: *"It just takes so long to sort and making sure it's all formulated right and it's all coded properly. And it's just annoying. It's very tedious ... And it doesn't actually provide you with the information you really need. ... It's not like the analysis portion where you actually run the numbers and come to conclusions. It's like literally just like making sure that everything is like coded exactly."* Like P8, P7 finds data cleaning a distraction from the kind of labor (here, analysis) that is more central to their role.

However, participants were more uniform when it came to important qualities in a data worker. Close to half of participants described the most important feature of a good data worker as attention to detail and/or interest in the datasets they worked with. P17 identifies the need to be detail oriented in data work: *"it's real easy to miss things."* P4 has a more elaborate description of the tension between needing to be detail oriented, but being confronted with the monotony of data work:

> *I think that anybody can really clean data or learn, you know, about how to analyze, properly analyze data. But, I think, if you're not interested in it, you know, you're probably going to get tired of cleaning it, tired of analyzing it because a lot of the job is just looking at the same data sets over and over again and trying to figure out how you can merge data sets or, you know, what those data sets can actually tell you. So, it's a lot of repetition. So, you've got to actually have some vested interest in what you're looking at.*

Collection and cleaning can intersect, too, to cause extra frustration when multiple data streams are involved. P1 describes: *"And I think the ability to see themes across*

the data. Right? Especially when we're getting it from all different sources, and to pick kind of how you're going to format it and kind of structure the data..."

However, the importance of good data stewardship takes on extra importance in light of the domain (public service work). P18 responds to a question about the most important quality of someone doing their job:

> *Integrity. I want to say something more technological, like you know, attention to detail, or ... like an ability to plan or something like that, but I think that when it comes down to the very end of things, when I took my stats class ... I realized that you can make numbers say pretty much anything you want them to say and I could easily like omit data from a dataset to make the story that we're trying to tell look more positive for us. But I think that in my heart of hearts, it's important to make sure that you're not doing anything unethical with the data, that you're presenting it as it is, letting people see the good but also letting them see the bad so that if there's a problem, they can fix it. If I hide it by you know, kind of selecting a certain subset that's only going to make us look good, then I'm doing disservice to the agency and G[*]d knows to the field at large.*

P6 stated similar sentiments: *"...From a government standpoint, I think the biggest quality is looking for accuracy... I think integrity's the biggest thing where you're working with, with data. I think the numbers are what they are."* Here, the position of P6's role—as a government employee—adds an additional weight to the need to both properly prepare and analyze data from a moral standpoint.

Summary: especially when our participants are the primary "data person" in their organization, they are solely responsible for multiple tasks in the data lifecycle, meaning they spend much more time on early steps like data collection and cleaning than other data workers (particularly those in high resources environments) who

may either purchase or obtain pre-cleaned datasets or be able to offload some of the monotonous or tedious steps of data cleaning. This is unique to this group of data workers, but the gist of their work—their perspective on the trials and tribulations of data work—demonstrates that they share a domain with their full participant peers.

3. *How data work gets done in the civic sector.*

How data work gets done in the civic sector

While our participants work in numerous organizations and have different day to day job duties, there are several themes that appear consistently in our interviews.

1. *Proximity to data.* The participants work with data that they are subject experts in, meaning that they understated the who/what/where of the data stream that results in the datasets they work with. In some cases, they themselves collect the data and engage in the analysis—e.g., P7's role involves developing policy decisions and they describes deploying a survey to relevant stakeholders. Similarly, P14 engages with offices and public service providers, such as the city code enforcement division, police and fire departments, and tax assessor's office, to negotiate data reporting procedures to collect the data the participant needs to answer questions of interest to their office.

   For other interviewees, their proximity to the context of the data's origination is facilitated by their previous experience either in the organization, or within the specific sub-sector of work, say, affordable housing organizations. P19 is an example of the first type of proximity. Having worked in GIS for several branches of the local government, including utilities and transportation, they know that the local utility companies—all of whom would have reasons to dig up section of road—lack a shared database. Therefore, a proposed moratorium system to prevent digging up portions of newly laid roads is not implementable until such a shared database exists

between these governmental and extra-governmental entities. As an example of the second kind of contextual experience: P12's job involves tracking use of their city's recreational facilities and the participant is in constant contact with facility managers. Together with one of the managers, they found that annual membership cards were resulting in visits to recreation centers were not being counted properly; this had direct implications for the amount of funding the center got.

Other interviewees rely on a data liaison for contextualization. The liaison is either directly involved with the collection process and can provide the data worker with more information, or their sole role might be to function as a go-between the individuals producing the dataset and the interviewee who performs analysis on it. An example of the first kind of liaison is the informal institutional knowledge that P8 and P9, who work at public utility service, source from their colleagues who work in the field (maintaining utility infrastructure):

> P8: *We've got quite a few gentlemen [field workers] who've lived there for two decades. They know our problem areas, so they'll say that's a problem area.*
>
> P9: *Yeah, they know our problem areas, so they'll say that's a problem area, it's back over where [old hospital building] used to be. 'Oh, yeah, I know exactly what it is.'*

Other participants have liaisons employed by their organization. P14, whose work involves budget analysis, describes how their liaison helps them track down points of confusion or concern: *"[The oversight office] they actually get the bills. And so I work with a liaison who, if I have a question on something, [the liaison] would work with me and the vendor, whoever it [of the contracting utility companies] may be, and be able to get an actual bill"*. Another participant—P7, who deals with civic infrastructure performance—describes how their data liaison (a specially hired

consultant) helps the participant make sense of unexpected data points in often-messy infrastructure datasets:

> *And the street address is that's associated with the [infrastructure landmark] but sometimes those are just made up addresses, depending on where the [landmark] is ... it's an address where if you look it up on Google Maps, it's just like a patch of grass or something... we send it to our consultant ... and then [the consultant] would just essentially go through and remove anything for which there's low confidence of accuracy.*

Here, P14's data liaison can make sense proxy data that would otherwise be considered "mislabeled" or even "incorrect". When P15 embarked on a housing data project and wanted to explore property tax assessments, they sourced their own liaison by bringing on a collaborator: an academic who specializes in the housing policy and was familiar with the data from their own work, underscoring the role of a personal network of collaborators. Another participant, P16, handles data for their city's public schools and explains how the school system has designated liaisons (roughly one per ten schools) to help interface between the tool builders (such as the participant) and the end users (teachers and administrators with questions). These liaisons, formally "specialists" for the information system, clarify data questions and manage feature requests that are passed back to the development team.

Summary: our participants have a unique proximity to the data they work with. Either they are directly involved in the dataset's collection, or they find ways manufacture proximity through a formal data liaison or ad-hoc interpersonal connections. This infrastructure allows them to capture the context of the data they work with and address anomalies or points of confusion in the data. Critically, for those removed from collection stage of the data they work, they might recognize they might misunder-

stand the data—despite their expertise in the domain—without the lack of (in)formal data liaisons who can restore that original context. The participants gladly welcome the data liaisons and consider them an important part of their workflow.

2. *Collaboration.*

> *"Collaboration hot; Separation old; We will give it all we've got; Together we have told!"*
>
> An idealized description of collaboration in data work, excerpted from "In the Data Kitchen" [314].

The importance of collaboration comes up repeatedly for participants. First, given the limited number of data workers within their respective organizations and offices, participants describe the role that collaboration with others has. P19, for example, describes how talking to other GIS workers throughout the city helps them and the rest of their team stay up to date on tools: *"Which is good because I need to keep fresh on the tools and the data...And then also be able to speak out to others with it."* Peer expertise provides support that P19 could not otherwise access. P6 illustrates how cooperation across organizations and offices both betters data flow and the projects involved:

> *Yeah. We were, they [offices sharing data] were here in person. Yeah. So, we were actually sharing that as we were gathering, they were gathering. I was looking at it but we would also go back to the departments that shared the different data. So, we checked them in the loop. That even made them a lot more excited about sharing it, even more. So, they were able to say, hey, what about this? Hey, what if we could layer this on top of that? ... So, what it began to do was it began to have some of the departments think a little bit more strategically in how they use the data be able to see it. Because I think one of the challenges is people gather the data and they*

go away and no one says what was that data used for? I have no idea. But when you go back to, continue to go back and show them the output of it. That's when you start to see some more creativity for the folks who are actually using it. So, that was kind of the case with that project."

For P6, not only does collaboration provide them with access to needed data flows, but it also engenders projects that integrate the expertise of the two departments.

There is another theme that P6 hits on: given the civic nature of the projects that the participants work on, they rarely work with a single data stream. Rather, most of the participants pool data from several sources. Particularly for those working in domain policy positions, they must collate data collected by different branches of the municipal government or various extra-governmental organizations. For example, P13 describes how developing housing policy requires the cooperation of several municipal agencies: *"So where we had to get the three agencies together and hash out all the details of how our programs work, how we measure things, and how it should be reported. And that has, it takes, it's kind of like you can't just have data people there. You have to [have the] domain person."*

The coalescing of multiple data streams that P13 describes has implications for inter-operability between datasets as well. As P4 states, *"... that's the biggest impediment I think in, you know, organizational partnerships... all our data's in different, you know, ... stages. Some of it's... still in paper format. I believe, if you start looking at nonprofits, you're going to see a lot more organizations that are still very paper heavy."* Not only can the nuance of the subject domain present hurdles, but so can the format the data is recorded and stored in. P1, too, describes the challenges of first obtaining data from another department and then attempting to work with it once received:

So, when I put out the call to each of the public agencies for the data, and I

basically just requested what they had, right? However, you're tracking it or producing a report, send it. And so, trying to be as flexible as possible... very slowly, we're seeing responses. And it all came in—I'd say it all came in, in Excel, but in very different levels of detail...Like some, like the [housing] stuff is super detailed, but it's also because of how they're required to put it into their system. And then the other stuff was like two columns ... it was like very different, and then everybody had slightly different fields. And it's like important information... I attempted you know, to put it into one Excel document, but I'm sure I [messed] some things up when I did that... I attempted then to start to merge it so that we could de-duplicate because a lot of it is the same... A lot of their money is in the same properties.

P1 is describing multiple public agencies providing funding for government-sponsored housing around the city; despite the overlap in properties, there exists no centralized system and P1 was responsible for matching relevant information about a given property from the records of the various agencies. However, when working with multiple offices or agencies within a municipal government, data doesn't necessarily exist in the same format given different standards (both prescribed by use and unintentional) as P13 describes: *"... the biggest problem is the data were in different formats, so police, if I recall correctly, used addresses and then fire had latitude and longitude."*

Participants frequently highlighted times that a lack of cooperation either halted their projects or presented major obstacles. Along with issues of data format, some organizations and municipal and state agencies fail to share data point blank. P13 their project lacked key data due to inter-office politics: *"We tried to get some data from the [relevant municipal office] and they weren't cooperative ... We had to leave it out."* P6 describes the problem as being endemic to the city in which they operate at an existential scale:

Well, in the city we when you say collaboration, first thing comes to mind: silos. Silos. I think the team that's here in this office, right here, are probably the most collaborative in the city... we cut across horizontal... any collaboration that takes place in the city, a lot of times, that relates to the data takes place with us. In terms of collaborating from department one collaborating with department two, there's not a whole lot of collaboration that takes place, I think. That's one of the areas I think city could benefit from.

P14 also describes the challenges of working in a state with an outdated data storage system and how this is compounded by the challenge of limited technical personnel:

It's just more time consuming maybe because we, you don't really have the resources as of yet to kind of just have the application like some states actually have it to where it's [on the] web. So they just go online. [A coworker] do[es] it and then it like sends it all the way to us. So everything is electronic ... [that coworkers] kind of gets it on one spreadsheet or you do that now but it's a little bit more automated?...[the same coworker] does it so how he has the application coded, he basically has the codes on the back end. And unfortunately, he likes to keep his stuff together so no one knows the codes and no one can actually unlock the spreadsheets to do it themselves.

Critical work in the state, as described by P4, is reliant on the personality and process of a single staff member.

Data work in the civic sector also requires collaboration and support from individuals both superior and subordinate. P12 describes the limitations of data collection when it requires inputs from individuals completing field work for their office:

We would love to be able to make it as where we can streamline it to where

when they go out they do everything digitally, but the only thing that would speed that up is being able to change everybody's job description where they'd have to be able to work. Because believe or not, we still have people who don't want to work with electronics...If it's not in their job description, you know, you'll find out real quick that they don't want to do it. And so that's part of the difficult thing with some of the people in the field...Like, you might have [an individual]who—and some of these metrics you'll see—whose job is to [perform infrastructure maintenance work]. You know, [this individual's] not going to want to be inputting [their] data [on electronic devices].

P12's attempts at datafication are hindered by a lack of corresponding enthusiasm for data by those in charge of the field workers, who have not incorporated extra allowances for data collection into their subordinate's schedules. The field workers are thus discouraged from engaging in data collection as it falls outside of the duties on which they will be evaluated and compensated.

Summary: collaboration—and the challenges it presents—are of elevated importance to participants. Unlike professional data scientists, they often need to communicate their findings to both those above and below them in institutional hierarchy. Having buy in from participants at all steps of the data lifecycle is critical for our participants; they often work on shared datasets that require extra technical and logistical support to obtain and use. Personal relationships between the individuals we interviewed and those in their professional networks are critical components of performing their role.

3. *Communication.* Communication is likely an important topic for any data worker, but our interviewees placed particular emphasis on its role in the civic sector. Many of our interviewees are responsible for formulating policy that directly impacts residents of the city, or they present their findings to elected officials who will take that action.

Hence, this group of data workers plays an important role in shaping how the contents of the datasets they work with come to be understood.

P5 describes specifically augment policy with data that will appeal to a board of elected municipal decision makers: *"We usually will do some research specifically on the district and its demographics and if we have them. I work on a lot of projects that involve construction in one way or another, on multiple assets, so I'll put together a map identifying them city-wide and then concentrations within districts and all that."* Here, P5 uses data as a way to personalize policy to a given decision-maker. P15 describes an incident where a higher-up incorrectly described a data dashboard to the organization's board of overseers, so the data workers in the organization removed the dashboard—which made information available to the public—as a result. We cannot share the actual quote in more detail without compromising the participant's identity. The accessibility of data here hinged on the behavior of a single individual higher in the professional hierarchy. Similarly, P17 describes trying to run intra-organization focus groups, which required the buy in of upper-level management: *"...they had trouble recruiting, and so I spent a lot of time trying to get leadership in the areas we were recruiting—these were operations employees, folks in operations—to really get the word out and encourage people to come to the focus groups and to work with the research firm. So it was a lot more hands on than I thought."* The project could not have continued without the support and cooperation of the participant's supervisors. P3 also describes how they had to seek buy-in from high-level officials to begin a project:

> *so we talked with the [Department of Education]... they actually do a really good job of providing data at the school level. We wanted it at the neighborhood level. So, we first discussed the need with the higher ups. So, you have to have some sort of access to even be able to have these conversations. And so, we've built up enough cache that we can go straight*

to the [second-in-command], the person in charge of data collection for all [the state] and say hey, can we come and talk to you about this.

However, unlike many organizations and organizational settings, our participants also require the work of colleagues in roles lower in the chain of command. P8 and P9 describe how they need to effectively communicate with the field workers in their organization, as well as city council members:

P8: *We do have conversations with them both directly and through their leadership as a conduit. Last February we had a series of special meetings on Saturdays to have a conversation about their frustrations ... So they called us in on a Saturday and we made our best attempt to show them... When we see 120 [infrastructure problems], this is what it looks like to us. I know to you it looks like chaos and, you know, you had one [occurrence] that it took three days to get resolved and another one, well, that was hardly a [big issue], it took like to an hour to clamp, but this is what it looks like to us. And also to say, hey, if you do a job and you don't close it out there, I have no idea that you did that. So when you're making the claim that you did this much work over the course of the day and I come back to you and say maybe, but I don't see that, and I've got hundreds of other work orders open, why haven't you done that, this is where we're missing each other... we had a slightly different format for trying to show them [the field works vs high level officials in the organization] that ... But admittedly, that's difficult ... We've got gentlemen and ladies ... who have no statistical background to be able to engage with you in a conversation about meetings or anything. It's just trying to communicate on this is what-matters-to-you information basis ... So that's a very different conversation.*

P9: *We focus mostly on the council members because the difficulty is—well, for the mayor's office, like for our whole agency monthly reports and we do share information on what's going on. They're actually invited to our [intra-organizational] meetings as well and do attend, so they can see kind of the nuts and bolts, but as far as local officials, as far as elected officials, generally we try to shy away from a lot of that kind of stuff, ... overall larger aspects, information about the system because... you get to a lot of situations where people might not understand, well, your area has a lot more development so you might see a lot more things in your area.*

The buy-in necessary to complete their roles was, for P8 and P9, dependent on correctly tailoring their work to two different audiences. P12, similarly, details delivering safety metrics to field workers in their organization in order to increase productivity after a workplace accident; we cannot describe the incident in more detail without compromising the participant's identity. P12's role dependent on the field workers continuing their work and therefore, they were stuck until the field workers resumed work.

Given that the civic sector often answers to the public living in its jurisdictions, our participants also describe challenges around making data available to the public. In some cases, the public is mostly unaware that the data can be accessed, as P14 describes:

I guess for us, I don't think people realize that our data is essentially available to anybody. Because we are a state entity, so we really don't own our data. The people own our data. So I think people don't realize that they can actually request a lot of our data from us. I think the issue sometimes is maybe a cost factor. You know, if it does require me to come

up with new data that we may already have in a different format, they do charge people... So I think if more people knew that, it will be, we probably have more like databases on more data kind of circulating around if people knew how easy it was. Or even that a lot of the data is on our website in spreadsheets.

P4 details the desire to make their office's data more available to the public, so that city residents can see what's actually being done: *"We also definitely want it to be more front-facing. So, to actually, you know, say what [a colleague] has said before, we want to be able to have something that's a front end where citizens can actually see what our office is doing in them and, you know, actually engage a little bit more. Because, we're really trying to up engage that with the community, for sure, which, you know, is why we engage in [an affordable housing] study."* Here, data dissemination is not only a component of P4's job, but an important part of their self-identified duties as a civic worker.

Summary: communication takes on an added weight for data work in the civic context. Participants describe a need to communicate both to decision makers and the general public who will be the recipients of the policies and data driven systems our participants design and/or implement. These goals share some overlap with those of data intermediaries, as we will discuss further in 3.2.4.

4. *Tools.* All of our participants explicitly reference working with Excel, though they do so to varying degrees; for example, P18 describes their Excel work as one of the more technical parts of their job and how they use *"buttons, macros, and user forms"* to augment basic spreadsheets. In contrast, P16 uses Excel only when they receive datasets in Excel's custom format. Beyond Excel, tool use is more varied between participants, owing to the diversity in precise role highlighted in Figure 3.1. For example, participants with more emphasis on backend systems referenced setting up

and managing Oracle and CRM datasets, while others more focused on collection described using tools like Google Forms and Survey Monkey.

However, a unifying theme for participants is that working in low or limited resource technical environments hinders their work. Specifically, their environments limit the tools they have access to and makes it difficult to obtain new ones. P6 succinctly explains why they are unable to acquire new tools when they find faults with (or major limitations to) existing ones: *"I mean, far as in the private sector, there'd probably be a yes to it [getting a better tool]. But in government, it's not yes."* Participants expand on the nuances of limited tool maintenance and acquisition in their work, describing the following:

- Limited funding. P7 describes how governmental workplaces rely on outdated software due to limited funding:

  > *I think the way we put our metrics into a manual process is very tedious. It's time-consuming. And then, that will go away when we automate it ... Now, I think probably for some of the departments, a lot of the information from the field is inputted into some of those systems, manually, as well, too. And I think if we had like more type of that, I mean if it was able to be sped up electronically, I think that would help out a whole lot... I think it's just a lot of different manual processes throughout government, in general. So, I think if I was looking at ways of improving government in this. Removing a lot of different manual processes that are, kind of, typical standard automated in a lot of different private sectors.*

  Here, there is a clear optimal solution that would better facilitate the work of P7 and their peers, but the solution is inaccessible due to a lack of budget. The same is the case for P16, who describes a similar experience working with a

local system that runs off a 25 year-old operating system version:

> *So it's more so kind of getting up to today's dates with software... it [the old operating system] has a lot of glitches... you have to like click stuff in order to go to the next page. So you would have to like click and then it pops up another one. You need to click that. Pop up another window. So it's really, it's really difficult to use. Most people won't even bother to learn how to use it. ... I mean, I had to learn how to use it because a lot of the information that I need is in there. And so, I learned how to use it but most people, they look at it and they're like, "I can't do it." Because if you, if you don't really use it all the time, you can forget and you won't be able to really get to where you need to get to because it's not really like you look at it and you're like, "Okay, I know how to click here." And like, okay, where do I put it?*

Again, there is a simple fix (updating the software system to a modern one), but there is not the funding to make that investment.

- High turnover. Another issue that plagues civic organizations is that of turnover. P4 describes how data initiatives spiral out of control in part due to the high turnover resulting from administration changes and how those initiatives are quickly abandoned:

> *Particularly for this department, just the huge amount of reporting we're now required to do. I think there's an interest in really showing what the [the city] and other agencies are doing... just for this office to accurately report on the information that we're required to with the frequency that we have to, we definitely need better formats for storing and recovering data. We can't, every time, there's a report have to scramble around for various spreadsheets that, you know, we might not even know exists because there has been some turnover in*

the office. And, when that happens, that information gets lost. So, it's important to have these structures where, you know, if somebody leaves, their information doesn't go with them.

Information that is collected is both rendered unusable by personnel turnover, as well as changes in priorities of elected officials. In some cases, after an initiative falls from priority, the data collection and reporting system is still actively staffed, even though the data is no longer needed. Sometimes the design of a system itself and the training provided to users, however, is responsible for a lack of use. P15 describes such a situation:

I think it's becoming a more like widely valued thing here, but we had sort of the traumatic experience—we got a new [customer resource management system (CRM)]. No one ever got trained on it, but it was mandated that everyone had to log in every day. I know we had some really bad data, like we still don't know go through [that CRM], like every department ended up going rogue and just either doing Excel or like we use our own project management software ... Every department does its own thing now, and we're, we're rolling out a new Salesforce program that's going to be pretty cool, but it's really hard to buy into that because like our organizational memory is like [non-existent].

Not only was the new system a waste of resources because it was never used, but it caused institutional fear of new software systems.

Subsequently, participant workflow is obstructed by a lack of reliable, affordable tools. Collaboration is hindered; P7 illustrates the ad-hoc nature of collaboration on a hand-coding effort with file type requirements:

[One team member] in particular who actually like created the macro that input all the data from the Google Sheets into the Excel, into much more

> *usable format. ... Basically I had an Excel file which all the data from the Google spreadsheet—well, the Google Survey dumped into a Google Sheet, which we then made into an Excel file in order to use the macros and then like organize all the data. Then we put it back onto a Google Sheet to like assign everything to everybody once the data was in a more usable format. But because of the macros that [the team member] had installed didn't work on Google Sheets a lot of the people had to redownload Google Sheets and give it to me in Excel format again once they were done with it. And I plugged it back in to Google Sheets. It was more complicated than it should've been.*

In very few workplaces would this be considered an ideal process. However, workflows in civic organizations are defined by what tools are available, not what is optimal. Reliance on outdated tools and systems also increases organizational security risks. This is particularly true for governmental offices and agencies, which are already the subject of cyber attacks and foreign interests [315]. One participant describes the process of manually updating changes to a large database while their computer system was held hostage in an organization-wide ransomware attack.

Summary: government offices and agencies, as well as independent non-governmental organizations, often have limited budgets that are reliant on public funding or private grants, rather than data workers' needs. Further, governmental entities are subject to high turnover stemming from frequent administration changes. Participants frequently mentioned the impacts of limited funding and high turnover on their work, noting that it required them to work with outdated and frustrating tools. Beyond worker experience, there are large institutional security concerns—not to mention citizen and customer data privacy—posed by these systems.

3.2.4 Discussion

Our analysis is framed by the questions Muller et al. present, as we seek to understand how peripheral data workers compare and contrast to their full participant peers and the subsequent implications for the future design of tools and processes. But more so than any specific implications for design, what emerges from our findings is the need to continually expand and refine our concept of data workers to include individuals on the periphery of the data work community of practice. While our participants comprise a particular subgroup of peripheral data workers—those in the civic sector—we theorize their concerns and practices are applicable to peripheral data workers in other sectors and professional settings. Namely, the civic sector is not the only one plagued by funding constraints and high personnel turnover; difficulties related to performing data work in this setting can thus be extrapolated. The one sector likely not touched by these constraints—the high tech industry—is already well-studied by members of this community, as discussed earlier in this work. Part of motivation for this work, in fact, is to identify, describe, and theorize about a broader range of data work beyond data science as it is commonly construed within and in relation to the high tech industry. Further, civic data workers, as an example of a peripheral group of data workers, may engage in practices that could be beneficially adopted by their full peers, such as data scientists.

Understanding data work as a community of practice

In order to address the question of who performs data work, we suggest that our conception of data work needs to be reformatted—as others have suggested—into a more inclusive community of practice. When we treat data work as a community of practice with multiple options for participation, we, as a research community, can begin to identify the breadth of individuals and groups contributing to and subsequently defining our data ecosystems. Until we do so, any efforts we make to understand the contextualization practices of data workers writ large will lack key participants in the various data ecosystems. Further, using

the community of practice framework, we can better ensure that the tools and systems we build are applicable for the full range of users; in other words, that we are developing tools that meet the needs of specific groups of users. We should, thus, pay attention to individuals on the periphery of data work who are full participants in their domain of expertise, and, similarly, understand data scientists as peripheral participants in the domains of the subjects they analyze.

Re: Who performs data work? We propose careful identification and examination of the workflows and processes of peripheral members of the data work community, who may be full members of another domain-specific community of practice. In order to understand who contributes to our data intensive systems, we suggest studying the needs and challenges of this group along with data scientists, data analysts, novices to the data work [297, 291], data wranglers [63, 263]. By arranging these subgroups of data workers in relation to one another, we can begin to understand common and unique experiences between them. Critically, we position the current conception of data scientist as full participants, with our interviewees as peripheral members, since, in many cases, they belong and identify with **another** community of practice as their primary affiliation. Similarly, data scientists with some expertise in a given subject area are peripheral members of that subject's community of practice. By broadening our understanding of how data work is completed across different fields, we can begin to make sense of how we, as a community, can both help transfer beneficial practices between different groups and also understand how data work is practiced outside of primarily technical organizations, such as large tech companies, and the difficulties therein.

In acknowledgment of a primary appointment or position in their domain of expertise, we actively avoid naming or labeling our interviewees as a class of data worker. Instead, we group them in relation to their peers who practice data work as their main professional practice. We have two reasons for this: first, our participants have their own professional identities that spans multiple domains. For this reason, we hesitate to label them as data

scientists directly, in acknowledgment of their expertise in their respective subject domains. Second, given the range of our interviewees and the variety between their day-to-day work and specific organizations, we risk minimizing the nuanced experiences of our individual participants. ”Peripheral data workers” serves not as a static classifier, but is meant to demonstrate relative position to a group already known to this community—full, professional data workers such as data scientists.

There is one immediate caveat to our proposition that data work can be understood as a community of practice that highlights an oversimplified view of the relationship between full and peripheral participants. In the traditional conception of community of practice [269], members on the outer rings of the community are often actively seeking to become more integrated into the community; for these participants, their role is much like that of an apprentice. When this apprenticeship process is well constructed, Lave and Wenger term it ”legitimate peripheral participation”, such that members on the periphery are engaging in meaningful sub-units of work in the domain under the supervision of a more experienced member, who both guides and legitimizes them. However, we do not mean to suggest that in all cases peripheral data workers *want* to become more full members of the community, nor does there often exist an explicit mentor-mentee relationship between the full and novice members of the data work community of practice. Besides acquiring more data skill that would help them with their present role, few indicated interest in becoming full time members of, say, the data science community, nor are do they describe being in frequent contact with full-time data work professionals. However, this distinction presents an opportunity for **collaboration** between primary subject matter professionals and primary data work professionals, given the extensive contextualization skills of the former and computing skills of the latter. Such a link would be more akin to a *partnership* than *mentor-mentee* relationship, because each individual would serve in both roles, as both a teacher (in their domain) and learner (in their partner’s). This bidirectional relationship represents a caveat to the work of full participants of the data work community of practice, we note, in that it

indicates that groups like data scientists may not have the necessary domain expertise to complete their tasks and actually rely on adjacent subject matter communities for assistance and contextual knowledge.

Given that contextualization requires extensive time spent in, or understanding, the ecosystem that produced that data, it is these domain experts—or peripheral data workers—who may provide insight about how that contextualization might be done. As they are close to the origins of the data, they may be best positioned to answer Loukissas' questions about the originating environment [85]. They can also make sense of factors that impact the data as it travels from collection to a summary or report handed to a decision maker. Having this knowledge makes them well-equipped to address contextualization throughout the creation *and* curation stages, as D'Ignazio and Klein discuss [47] and bring us closer to Monroe-White's concept of *emancipatory data science*, which supports the conscientious and consent-concerned use of data created by marginalized communities [276].

Understanding the practices of peripheral data workers

Muller et al. [74] pose the questions of how data workers go about their tasks, how they collaborate, what the implications of their work is, and what their needs regarding collaborative tools are. We answer these questions in turn, with our responses focused on the group of peripheral data workers identified earlier in this work, who belong to an understudied demographic of data workers.

Re: How is data work performed? As discussed earlier in this work, most of our knowledge about the way data work happens uses data scientists as the study population. While data scientists are an important and sizable community of data workers, they receive an outsized amount of attention in comparison to groups more on the periphery. In particular, many efforts to address bias and fairness in algorithmic systems involve warning labels—or, said more mundanely, a kind of nutrition label—about who and what is represented in that dataset, along with clear statements about the situation and context from

which the data arose. The solutions are geared towards the community of highly technically skilled data scientists, without specifying how data scientists should source this information, or practice conscientious consumption of it. Meanwhile, many peripheral data workers work in immediately public-facing setting and have found ways to meaningfully practice these conscientious consumption and reflection practices. Our interviewees are a good example of this: the data work they engage in has a direct effect on city residents and various vulnerable populations (those served by the non-governmental organizations). As shared in our findings, our interviewees have ample access to the data they work with, either having collected it themselves or via a data liaison. If we want to implement documentation procedures to address issues of representation in datasets, these data workers are uniquely positioned to show us how those practices might be established in other groups within the data work community. Further, given their public facing role—or implications of their work—our participants share some overlap in characterization with ”data intermediaries” (DIs), or individuals with data skills who connect members of the interested public, particularly members of marginalized communities, with data related to their interests or concerns [281]. However, their are two notable differences between our participants and DIs: first, the latter do so as avocation while our participants are salaried employees of public-facing organizations and institutions, and, second, our participants are more valuable given their level of contextual knowledge, while DIs are sought as partners for their technical skills.

Re: How does collaboration happen in data work? This group of peripheral data workers have developed systems, both formal and ad-hoc, to answer their questions about a given dataset. Consider the aforementioned data liaisons, as well as the institutional knowledge that P8 and P9 rely on to make sense of errors in their dataset, respectively. This corpus of interviews also highlights the crucial role of cross-hierarchy communication; ideally all data workers could be able to communicate the contents of, and concerns about, a dataset they work with to both those in their technical setting, but also to the laypeople

who are affected by those systems. As there are increasing concerns about the fairness of algorithmic decision making in all facets of life, it is worth noting that our interviewees, all of whom work in the civic sector, play a part in systems that city residents usually have no option to disengage with. For example, city residents are limited in their choice of municipal services and utilities. Those who receive aid from the various non-governmental organizations represented amongst our participants very rarely enter the partnership in a position of power, which would enable them to ask how and why decisions were made and what their data was used for. In the more general case, it is important for citizens to be able to understand the open data the respective governments and organizations seek to publicize; civic data may be open for access, but, as Boehner & DiSalvo describe, closed by lack of access to context [283]. As a community, we should aim to both recreate and further extend these contextualization and cross-hierarchy communication practices.

Re: How can we design tools for collaboration? Beginning the work we found that discussions of both data science and civic data often suggest that idealized contexts and practices would be seamless, with a free flow of interoperable data across a common toolchain and coherent process [316, 317, 318, 319]. Whether or not that is desirable or achievable, it is not the case. We found these peripheral data contexts and practices are seamful: characterized by misalignments, inoperability, and the ongoing labor of translation and negotiation [320]. As Inman and Ribes argue, neither seamlessness nor seamfulness should be considered inherently good or bad [321]. Seams mark strategies of revealing and concealing how a system or process works. As such, seamlessness and seamfulness afford different means of access and control. Much of the labor of these peripheral data workers is tactical, developing work-arounds for the seams in their context and practices. In some cases these seams are particular to the data, e.g. P8 & P9 utilizing their personal experience (and that of the field technicians) to correct faulty addresses. In other cases, these seams relate to tools—consider P7 and their coworkers' issues with analysis software, requiring switches between Google Sheets and Excel—and in still other cases these seam

are socio-technical, e.g., P12 struggling to get their field technicians to input data when it seemed like an unnecessary imposition without payoff. These practices echo the labor of work with seams that Vertesi describes in her ethnography of space science. As Vertesi describes, even well-funded scientific endeavors—such as sending a robot to mars—require negotiating, or what she refers to as suturing, seams. It is not surprising, then, that such seams and tactics of suturing exist in domains with fewer resources, such as local government and civil society. These seams and the tactics to suture them become opportunities for joining and expressing data in novel and useful ways. For instance, Dailey and Starbird have studied and described how residents, government workers, and journalists sutured the seams of social media platforms to create needed and responsive communication channels in response to a natural disaster [322]. We argue that describing and attending to such resourcefulness in seamful data work can lead to both a better appreciation of this labor and inform the design for data tools and processes for domain experts.

Designing data work tools for domain experts

The answer to the question of tools needed for data workers should be the result of analysis of different constituent groups in the community of practice. Our findings illustrate several challenges faced by civic data workers, as subject domain experts, but peripheral data workers. Namely, there is often a lack of funding, or funding is provided by mercurial grants. This results in limitations on what new tools can be acquired and an over reliance on legacy systems that produce inflexible data streams. For those of our interviewees who learned their data work skills on the job, they had less exposure to broader groups of tools and tend to stick to those tools with which they were already familiar, even in cases where they would like new ones. For those who work for local and state government, there are also issues with documentation and archival practices, as new administrations present new data collection priorities and practices and personnel turnover is high. As local and state governments are also increasingly targeted by ransomware attacks, there are also specific

security concerns both for tools used and data storage methods.

However, these limitations are likely not unique to the civic sector. Other data workers on the periphery are also embedded in organizations with non-technical foci; many of data workers potentially experience the same obstructions. Critically, these issues are much less likely to bother data scientists, who often operate within large technical organizations. In contrast, peripheral data workers need tools that 1) create flexible, more interoperable, and sustainable data streams and 2) provide accessible entry points for users with less technical backgrounds. In building these tools, we should incorporate space for the contextualization these data workers practice, both concretizing their observations and sensemaking practices, as well as facilitating documentation in roles with naturally high-turnover (such as those occupied by political appointees). By limiting our conception of who performs data work, we ignore a specific and documented need for data work tools that meet these requirements; this is something our community can and should address in our tool designs and user experience assessments.

Re: Implications for stakeholders and affected population? In designing these tools there is also the opportunity to bake in some of the contextualization or reflection practices already suggested by this community. If we understand contextualization as one type of collaboration (namely, conjoining context and the textual data), some of these guides can also be adapted to the specific working context of peripheral data workers. For example, what would it look like to build data analysis tools for peripheral data workers that helped them concretize those contextualization practices which they already engage in, and also add new ones? We can imagine something like space for Gebru et al.'s datasheets for datasets [288] made in these tools, that helps convey context between collaborators. The datasheets could be extended, too, to the civic data work space. Data workers could be asked to record not only recommended uses, but to explicitly warn what imagined uses would be inappropriate or reason for concern, given the naturally transient staffing of governmental organizations and the likelihood that a given dataset may be reused or revisited

in the future when those most familiar with it are not around.

Re: Tool Evaluations. While we do not directly answer how to shape evaluations of tool learning-curves and usability, our findings suggest some preliminary ideas. First, tool uptake is significantly shaped by institutional resources, meaning that evaluations must take place in situ and in a variety of environments and account for the price of access to the tool. Second, tools should be graded for their usefulness—once learned—for given application areas, to help individual data workers in low-resource environments decide if they are worth the payoff. Third, tool evaluations should be made available and accessible to non-researchers, to provide a centralized place for peripheral community members to search for new tools.

Limitations

We note that all of our interviewees within this sample work in the civic sector in the United States of America. This choice was intentional for two reasons. First, while language around, and specifics of, municipal services varies by community and national setting, common concepts such as water treatment (or lack thereof) are more accessible than a specialized discipline, such as a natural science. This allows us to focus on the way in which data is handled by our participants, rather than the background of their work. Second, within the United States, civic organizations—either explicitly governmental or not—are typically sites of limited financial resources. However, we acknowledge that findings from this sample population describe technical challenges faced by a one group of individuals working in an applied field of data work and is limited, too, by the social and cultural aspects related to the singular nation of focus. However, we believe that their technical challenges, or balancing a need for better tools with a limited budget and high turnover, is not unique to the civic sector (or the United States) and can be found in many organizations, particularly those that are not large tech companies, where only a small portion of the population of all data workers are employed.

Finally, our interviewees' demographics, as noted in the results section, also vary some from national averages in computing and information technology. While many of our participants are not employed in roles that would necessarily be counted in these categories, they are the categories that many full participants of the data work community would be counted in. We theorize that while our survey population may not thus be representative of the computing and information technology workforce, it may be more accurately representative of the demographics of those on the periphery of data work, who are spread out across multiple fields, and may have slightly lessened gender and racial disparities compared to tech fields. Given that our participants worked in various organization types, this is difficult to state with certainty.

3.2.5 Conclusion

In this work we interrogate data work as a community of practice, using Lave & Wenger's framework [269]. We identify the unusual role of data workers on the periphery of the community of practice. They are often full professionals in a different community of practice but engage in more robust efforts to contextualize the data they work with, as compared with more full practitioners of data work (e.g., data scientists). We report the findings of our 19 interviews with civic employees in large cities in the United States who engage with data, but who are peripheral members of the larger data work community. Pulling from these interviews, we suggest how other data work practitioners could learn from these contextualization methods. Further, we aim to raise the profile of data workers on the periphery and suggest that as a research community, we consider the unique challenges of performing data work in low resource technical environments. Our goal is to contribute to both better contextualization practices and systems for full data workers and promote the needs and concerns of data workers beyond data scientists. Our findings address open questions in the burgeoning human-centered data science community [74], including highlighting a wider range of data workers, along with their particular data work methods and workflows,

along with their role in the greater data work community.

3.3 Chapter Conclusion

In this chapter, I traced the history of AI and data science back to cybernetics, with its footing firmly in the Libertarian vein of the 1960s Counterculture movement. What this means in terms of the problems we see today is that those same issues of obliviousness to nuance are likely the newest incantation of an early view of computing that prioritized universality of local context. In particular, the mindset that took hold among the early generation of developers of computing for the masses was predicated on a belief that computing was the ends, rather than a means to the end. The host of sociopolitical problems then at hand were to be addressed with a liberation of information (via the earliest incantations of what we now know as the World Wide Web) that would link and unify the global population. Instead, this predilection for a single story of culture and history generated a technology industry that, as I elaborate in the first section of this chapter, eschewed the perspective of individuals, particularly towards the bottom of the corporate ladder, for the ideologues at the top.

I return here to the notion of *cyberfeminism*, as introduced earlier in this work (1.3.1)—which was formed as a pluralistic, situated critique and response to cybernetics—as an alternative approach. What if our contemporary practices of AI (and data science before it) had been premised instead on employing data to tell partial, contextualized accounts of events? What if we understood datasets as yielding an incomplete story? Where would that put us, instead the endless labyrinth of discriminatory (and usually binary) systems in which we currently exist and in which we are continually confronted with the failure of such systems, usually along some sensitive social axis?

As an exploration of what an alternative model could look like, I turn to the case study of civic and non-profit domain experts-as-data workers in the second half of this chapter. Here, we see a group of data work practitioners who are concerned with understanding

the data narrative as a means of deeper insight about the communities they work with. The robust contextualization and documentation practices they employ—ranging from reaching out to knowledgeable parties, sometimes at their own expense to tracking down deeper readings from a host of intra- and extra-organizational colleagues—help them understand the stories that the dataset puts forth. I offer these contextualization practices as a foil to the paradigm more typical in contemporary commercial practice; instead of mitigating the individual perspective (as "bias") in industry-grade practice, the data professionals in these civic and non-profit roles are focused tracking down and embracing as many relevant perspectives as they can find.

The fact that civic and non-profit data workers can manage these practices despite what are usually relatively small operating budgets, outdated computing systems, and limited staffing suggests that commercial enterprise should have no major obstacles to doing the same. The practices demonstrated by the civic and non-profit data workers is more akin to a cyberfeminist approach to performing later data analysis and even system development from this data; each dataset is understood as a social product that requires understanding the perspectives of the individuals who shaped it. In many cases, by the time the data workers we studied begin working on these datasets (if they were collected by someone else in the organization, for example) they were acting as dataset auditors, too. Their observations and questions about different values, for which they reached out to potentially knowledgeable parties, contextualized later use and understanding of the dataset.

The work that follows in the rest of this dissertation is inspired by these practices. Specifically, I focus on the purpose-driven examination of datasets, focusing on benefice to human health and happiness. This includes how the data workers we studied built out their contextualization methods and their robust collaboration practices. In the chapter that immediately follows this one, I tell the story of critical data literacy at DataWorks. The critical data literacy curriculum centers around understanding datasets as sociotechnical products that are transient, incomplete, and messy records of events and plans. At Data-

Works, once the Data Fellows were equipped with critical data literacy, they could convey their lived experiences and perspectives into the bigger picture of their work—-namely, in making sense of the AI and ML systems to which they were contributing.

CHAPTER 4

WHAT PERSPECTIVE BRINGS TO DATA WORK

In the prior chapter, I examined the history of AI and data science, tracing them back to the tradition of cybernetics. The particular strain of cybernetics practiced in the U.S., which went to on to lay the groundwork for the contemporary American computing and software industries. This particular strain was one with a heavy Libertarian leaning, in which computing was supposed to serve as both a universalizer and an equalizer, to eliminate the need for governments and borders; in other words, it was (and continues to be) a phenomenally technosolutionist vision.

As a foil to cybernetics, I question whether our current AI futures, homogeneous and unappetizing as they are, would look the same if *cyberfeminism*, instead of cybernetics underpinned the computing ethos? Where cybernetics calls for collapse of detail, into a system of discrete command and control units, cyberfeminism has an insatiable appetite for the pluralist, infinite visions of the individual, where the individual could be the user, the developer, or anyone in between. Each individual carries their own perspective along with their experience and a system designed to incorporate and respect each would look fascinatingly dissimilar to that with which we are confronted today.

I then introduced work pertaining to the dataset contextualization and documentation practices of a series of civic and non-profit workers, as a demonstration of something much closer to a cyberfeminist practice. Now, in this chapter, I explore what it means to try to re-create that practice. My investigation is centered at DataWorks and the first half of this chapter is devoted to the development of a novel critical data literacy curriculum. In tandem with the introduction of that curriculum, I offer a case study in data contestation and refusal that honors a cyberfeminist mentality. Specifically, in practicing their critical data literacy skills to convey their lived experiences and perspectives to the client in question,

the Data Fellows managed to reshape a data task (and resultant system) both conceptually and materially.

The last portion of this chapter is reserved for discussion of how we actually record, archive, and share the resultant meta observations (or, *paradata*) of dataset creation and development. I introduce a Google Sheets add-on, *Datum Fieldnotes*, that helps the data workers I've thus far described expedite their process of data review and auditing. Datum Fieldnotes automates additional documentation, to both support data workers who are learning and testing out new data transformation functions as well as create an interrogatable record for the purposes of paradata examination and excavation. While this tool is a follow up to work describe in the prior chapter (3.2), it is intended to serve as a documentation and reflection tool for scenarios exactly as the one described in the case study in section 4.1. Further, the *Datum Fieldnotes* allows users to augment a dataset (contained in a Google Sheets workbook) with per-datum notes that can be used to flag interesting or discussion-worthy data points, or simply flag a data point for further investigation.

4.1 Designing for Critical Data Literacy

As DataWorks took on clients, we saw an opportunity in conducting work on datasets needed to develop and train artificial intelligence (AI), machine learning (ML), natural language processing (NLP), and computer vision (CV) models. The work of labeling, annotating, and standardizing data sets is tedious, yet a skill that can be learned by people new to technical work, which makes it apt for DataWorks. In addition, we believed there were situations when the knowledge and perspectives of the workers could be useful in processing the data: situations when they would bring a view that would add value to the data. We reached out to researchers at multiple institutions, describing DataWorks and promoting our services, and several projects quickly materialized, including the Children's Book Project discussed earlier in this proposal (2.2). One such project involved clips of infants lying on their backs and labeling when they kicked their legs. This was used to investi-

gate a system for early detection and intervention for infants at risk for motor impairment. Another involved labeling clips of video game play to train an algorithm to identify pairings between player strategies and game mechanics. After the first few such projects, we developed a workshop series to familiarize the workers with basic ML concepts and mechanisms, combining our expertise in participatory methods and the learning sciences. These workshop activities did not require prior experience in programming or computer science. They were intentionally designed as non-computational approaches to understanding ML that presented the larger concepts in the context of the everyday. These activities are designed as scaffolds, or supports for students to move from what they know to tackling new knowledge [323]. In particular, these activities scaffolded an understanding of ML among novice data workers by connecting to their work experiences and current affairs, that is, their individual and collective lived experiences. The purpose of this workshop series was thus twofold: to contribute to the workers' contextual understanding of data and data work and to help the workers screen prospective client projects and assist them with refining their goals for the data task. In other words, we intended the workshop series to provide both critical data literacy skills and contribute to their ongoing technical and organizational skill development. These activities, and the outcome of a critical consciousness towards data and ML that resulted from these activities, affected subsequent DataWorks projects and processes in profound ways.

The workers began the workshops with an acute awareness of systemic racism from their lived experience. As Black citizens in the United States, they face racism every day. They are well aware of not only experiences of racism such as slurs, but also of how White people's perceptions of them affect their ability to get jobs and their treatment in the workplace. They are also well aware of how, for some of the workers, their neighborhoods were historically and consistently provided less services, from bus routes to grocery stores. We share this because we want to be clear that we are not claiming that the workers had no critical consciousness before these workshops. On the contrary, racism was ordinary in

their lives [63, 324, 325]. Instead, our attention is on a critical consciousness about data, algorithms, and, more generally, computational systems and technologies. Building off the critical consciousness and lived experience workers already held in different domains, the goal of this workshop was to help workers analyze potential uses of data-intensive systems in conjunction with their role on those projects. This intentional effort to promote worker agency is one element of DataWorks that sets it apart from other data work sites. Critical perspectives on data were not as familiar to the workers before the workshop. This is unsurprising. Until recently, discussions of bias in data and algorithms were primarily academic, with scant coverage in popular media.

The workshop series took place once a week over approximately three and a half months. Each week, the workers and a researcher would meet for about 90 minutes. We built each workshop activity around a critical data literacy concept and a data set or algorithm through which the workers would explore that concept in practice. By anchoring the concepts in data and using the techniques of data analysis and operations that were part and parcel of their client work, we wanted to materially and experientially demonstrate the connections between data, practices of data work, and issues of justice. The workshop was separated roughly into two halves; the first half focused more on understanding ML, while the second examined basic data skills in conjunction with issues of equity and justice.

Through this, we expected these workshop activities would contribute to a critical consciousness about data, within which the workers would come to recognize the potentials, consequences, and limits of data work. For example, the first activity of the workshop series was a discussion of evidence-based assertions, that is, how data can be used to advance a narrative or support a position. The key to this was to develop an awareness that data is not "objective," but can be used, and even distorted, to present a specific narrative. The case study for this example was a map of the COVID-19 cases in the Southeastern United States (where DataWorks is located). When the map and accompanying data set was introduced, we asked the data workers to spot what was "wrong" or "off" about the graphic.

After several minutes and some nudging, they spotted that the scale of the graphics made no sense — the same color scale (used for two identical maps, to showcase progression over time) was matched to different case values. The workers then explored a dataset about COVID cases in neighboring states. The researcher then posed a series of questions. The questions were formed around understanding what arguments, or narratives, might be developed from the data. These included questions such as, *Could the data be used to make the argument that COVID-19 affects women more than men?* and *Are people with existing comorbidities at a higher risk of death?* and *Could the data be used to make the argument that people in Southern states are unhealthier than in other parts of the country?* These questions were designed to demonstrate the complexity of situations represented by data, while also drawing on workers' lived experiences as Southerners in the COVID-19 pandemic. For example, the question about Southern states, which was depicted as having a higher case load, overlooks the context of those states, which are generally a more rural and underfunded part of the United States that, in some places, lack in both basic preventative and emergency healthcare infrastructure. Here, the objective was to demonstrate how data can be used to further an agenda, and how a critical examination of what the data shows can help uncover con-contaminate social and technical phenomena. The researcher also asked reflective questions that prompted considering the limits of this data set, such as *What claims would you feel comfortable making based on this dataset? What information about the dataset would you want to have, but don't?* and *What reservations would you have about working with this dataset?* After answering these questions individually, they were discussed as a group. Through that, they continued to collaboratively articulate a collective understanding of this particular data set and develop critical perspectives on data more generally.

More technical terms and processes were introduced as the weeks progressed. At the same time, we also introduced more social and political connections between data, ML, and everyday life, by drawing from news stories and integrating videos from organizations

such as Data for Black Lives that directed our attention to multifaceted aspects of data: from being responsible stewards of data, to the weaponization of data and algorithms, to the pro-social potential of data and data work. The workers were less familiar with these situations, particularly regarding data. For example, while they had experienced racism in job hiring, they had not previously considered how data and algorithms might perpetuate and amplify racist hiring practices. Once realized, the implications were unsurprising to them, as yet another experience of racism.

Throughout, we designed the weekly activities to support worker agency in developing critical data literacy. To this end, we believed it was not enough to make workers aware of the themes of critical data literacy and the potential harms of data and algorithms. It was also necessary to assist the workers in developing skills that they could use to address harms caused by data or algorithms and take direct action through their work. For example, an activity on ML applications included developing a list of questions to ask whenever starting work with a new dataset about its contents and potential applications. Workers practiced this list on hypothetical scenarios like a medical company that develops a system to diagnose different kinds of bug bites without having patients interact with a doctor in person; instead, patients submit photos of their bug bites for automated classification via an app. Potential concerns for this scenario discussed in the workshop include issues of medical ethics (removing in-person care), privacy (storage and processing of user photos), and fairness (whether the system has been trained to recognize potential variance of bites on different skin tones), among others. We reiterated the real-world implications of these projects; for example, the following week after the bug bite recognition system, we began with a discussion of a news story that had emerged in the interim, in which Google was launching a dermatology AI app in the European Union.

One outcome of this workshop series is that workers expanded their vocabulary for talking about data, data work, and issues of justice. That is, they developed a critical consciousness and capacities to express that critical consciousness. For example, one work-

shop focused on the task-work systems often used in data labeling projects, like projects they were doing. At the start of the workshop, the workers recalled earlier experiences with such systems, saying they had not always been compensated and that the work seemed unfair. After the workshop activities and discussions, the workers could speak about these systems and issues with critical insight. As one worker expressed in a written reflection referring to a task work platform, *"This system is set up to exploit workers because there is a low level of transparency and virtually no workers' rights. There should be regulations in place to protect [crowd] workers."* Such a reflection reveals an astute understanding of both the systems and practices of data work and potential courses of action to address problematic conditions. Significantly for relating critical consciousness to workplace democracy, the worker casts this problem not merely as a technical issue, but also as a labor rights issue. Seeing the connection between data, their work, and their experience of oppression was overwhelming for one of the workers. After viewing a Data for Black Lives video, she left the room in tears. A while later, she returned and stated that she was overwhelmed, and that it was *"us against them"*, and saw herself as a data worker in a position to act against the systemic racism and oppression against herself and her community.

4.2 Case Study: Expressing Perspective on Data in the Workplace

4.2.1 Expressing critical consciousness of ML in a data annotation project

A few months after completing the critical data literacy and ML workshops, a computing researcher from another institution approached us about a labeling project involving African American Vernacular English (AAVE), which is also referred to as Black Vernacular English. The client used the term AAVE, so we are using that term in describing the project. After discussion, we thought this project might be a significant opportunity for the workers, who at that time were all Black Americans, to engage in a project that built upon their positionality and demonstrated the value of their expertise. From a design justice stance of "nothing about us without us," such a project was an opportunity for the work-

ers to contribute to shaping technology that was about them and would affect them [326]. Following our usual process, we agreed to an exploratory foray into the data set and accompanying work led by the workers. Very quickly, the project revealed complexities and potentials of supporting critical consciousness in the workplace as the workers questioned, resisted, and in due course, re-negotiated the project's design and execution in meaningful ways. What had been hypothetical scenarios in the prior workshops materialized as real-world events as the workers, researchers, and clients engaged the issues of bias and harm in data sets and data work.

The project was to examine short statements and annotate the statements as examples of AAVE or not. Some statements were taken from social media channels, such as Twitter or online forums, such as Quora. Other statements had been algorithmically generated based on a model of AAVE. In other words, some of these statements were not expressions of AAVE authored by people but algorithmically generated AAVE statements. The purpose of the task was to assess the capacity of the algorithm to generate AAVE, knowing already that the algorithm was flawed. From that assessment, the client believed they could modify the algorithm. The client sought out DataWorks specifically because they wanted to approach this project responsibly and ethically. Rather than distributing the work through an online crowdwork platform, they wanted to use workers who were culturally situated for the annotation task to produce data, which would, ostensibly, enable them to create a more accurate and appropriate computational model of AAVE.

Numerous problems quickly arose. To begin with, the workers were uncomfortable with the notion of a singular African-American vernacular. They pointed out that their experience of a Southern vernacular was distinct from what might be a vernacular in Chicago, Los Angeles, or Boston. If they were experts, it was only regarding a Southern Black vernacular. Then there were issues with the data itself. Both the algorithmically generated phrases and those taken from social media and online forums were often offensive, leading to discomfort among the workers when annotating the phrases. Most of the offensive mate-

rial was sexual in nature, some of it was homophobic. Some of it was also racist, in particular expressing anti-Asian and anti-Asian-American sentiments. The workers immediately identified these phrases, and the condition of having to work with them, as inappropriate and unacceptable. They requested that the manager reach out to the the client, make the client aware of the issue, and ask for a remedy. That recognition and request, as simple as it might seem, is an important expression of workplace democracy, as the workers had the capacity to identify work conditions as inappropriate and the agency to request those conditions be addressed. This is in contrast to the dynamic at many data work sites, such as content moderation firms, where workers are neither encouraged nor supported in voicing concerns about their work [295].

In addition to the content of the data being offensive, the workers began to find the very concept of the project problematic. In daily check-ins and conversations, the workers began to question the premise behind the project and the project's actual outcomes. They raised concerns such as *"Who really wants to be able to identify Black speakers?"* and *"Once someone has been identified as a Black speaker, then what?"* These questions are similar in kind and subject to the questions raised in the prior workshop series. As these questions emerged, the workers seemed to be applying the critical perspectives from those previous conjectural activities to this client work. As the questions and conversations continued, several of the workers arrived at a specific concern: *"It seems like this could be misused in a way that will hurt my community."* And then, they requested a meeting with the client. In the meantime, we collectively agreed to put the work on hold until they had expressed their concerns and the client had adequately addressed those concerns to the workers.

We reached out to the client and the first meeting went poorly. The graduate student who was leading the work met with the workers. He was not a person of color and demonstrated little experience talking about race. He listened to the workers and told them about the purpose of the study: to demonstrate that current algorithms for AAVE were inadequate and that their work would help substantiate that claim. However, that explanation, and its

presentation, did not build confidence in the workers or adequately address their concerns. Two Black American Ph.D. students also attended the meeting—we thought to help alleviate power imbalances. This proved unnecessary, as the workers directly, and without hesitation, expressed their concerns with the data work. The workers made multiple suggestions, including that a Black researcher would have better insights and sensibilities to conduct this work, and that rather than reviewing an algorithm already known to be flawed, it might be more productive to consider how to more directly generate AAVE better. The graduate student was flustered and returned to the lab to share these concerns with the lead investigator. In the meantime, the workers requested that the project remain on hold until their concerns were addressed.

Several weeks later, the graduate student met again with DataWorks, with three changes to the project, each in an attempt to address the concerns raised by the workers. The first was to use a filter to remove profanity from the data set before sending it to DataWorks, to address the problem that the algorithmically generated statements were at times offensive. The second was to change the response form to include more nuance besides correct or incorrect. Workers could instead express if a translation was correct, incorrect, or concerning. If a translation was marked as concerning, the research team would follow-up to understand why it was labeled as such. Finally, an open entry field was added to the form used to process each phrase, to enable the workers to provide ideas about why the translation was wrong, and express other concerns. With this new responsibility, what had previously been a task defined and dictated by researchers had, in effect, become a co-creation task between the data workers and the researchers. It is important to note that during this process, the workers continued to be paid their regular hourly wage and were encouraged to discuss their concerns while ”on the clock.” The workers were compensated not only for performing the annotation work, they were employed and accorded due respect as consultants on the project, even during the stoppage of the annotation work.

With these changes in place, the workers resumed the task and completed the first batch

of labeling. However, several months later, the student researcher sent over the second batch of AAVE translations for labeling. New workers were on staff who were concerned with what they heard from the veteran data workers about the previous batch of data. In a staff meeting, the data workers expressed their concerns after taking a pass at the new data set. The data still contained profanity and inappropriate content. Moreover, the new workers had no context for how this work could be helpful to their community–they only could see how it would be harmful to the Black community. One of the DataWorks research staff, a faculty member, reached out to the lead investigator on the project and advisor to the graduate student involved. They talked about the data workers' concerns. They brainstormed how to implement the suggestions previously made by the workers, including changing the lead student researcher to a Black Ph.D. student and focusing on translations rather than annotations based on the workers' previous feedback. Subsequently, the lead investigator met with the workers. She answered the workers' questions about the study's goals, assuring them that it sought to create technology inclusive of dialects and vernaculars beyond Standard American English. She related how other dialects were also part of the study, including language variants associated with her positionality as a person of color and an immigrant. She understood their concerns, and while she was not Black, she could empathize with how language is used to advance racist perspectives. And she sought out compromises.

The first was to review the algorithmically generated phrases for offensive keywords with a stricter protocol and delete offensive entries from the data set—in effect, to do what the workers had asked for already but which the client had failed to actually follow through on. The second was to ask the data workers for translations in addition to annotations. Notably, the data workers rejected this idea; they felt it would produce biased data as they knew only a Southern dialect of AAVE, and they *"could not represent all Black people."* The third was to change the primary graduate student investigator to a Black student. One of the new data workers was insulted by this suggestion. They told us it signified their

grievances were about the race of the researcher rather than the data and the purpose of the data. However, the workers who had previously requested this and researchers explained the prior reasoning for this, and subsequently followed through with this change. Finally, the lead investigator suggested a data use agreement that the workers would write. This data use agreement would state the intended uses of the data and what the data could not be used for. Once the data workers wrote the agreement, the lead investigator would sign the use agreement and commit to it. Only after the offensive data was removed and the data use agreement was signed did the workers continue and complete the project.

Of course, data use agreements, unless bound to policy, regulation, or law, are nebulous and nearly impossible to enforce [327]. This data use agreement depends upon a social contract and goodwill. While that is problematic, we should not dismiss the importance of the worker agency in this situation and the critical consciousness that developed. It is important that they identified consequences of this data work. It is crucial that they were able to act upon that, to choose to stop the project, to require accountability from others, and then to determine under what conditions they were willing to act. This is essential to fair and just approaches to data work.

4.2.2 Discussion

The workshop series and subsequent AAVE project demonstrate an emergent critical consciousness and provide insight into a workplace democracy. In the following sections we reflect upon this, connecting back and building upon the research literature and themes that motivate this work. First, we name two characteristics of workplace democracy that other researchers and practitioners can use in their interpretations of the workplace: refusal and agonism. Then we discuss the difficulty and importance of moving between designing work systems, designing organizational frameworks for action, and designing the industrial relations context. Finally, we reflect upon the institutional labor of workplace democracy, how the positionalities of the workers, clients, and researchers affected this project, and the

changes we have made to DataWorks based on this experience.

Refusal and Agonism in the Workplace

As the workers developed more understanding of the implications of data, they approached data work with increasing skepticism. In the AAVE project, upon realizing it was dangerous to their community they chose to halt the work. This caused a disruption to the flow of work within DataWorks and prompted challenging engagements with the client. **Such trouble is appropriate and necessary.** Critical consciousness should be disruptive to the status quo. Such trouble is undoubtedly what was intended in the formulation of critical consciousness by Friere [328]. Critical consciousness finds its apogee in action, and action against an oppressive status quo, intentional or not, will be disruptive. In the workplace, for that disruption to be meaningful, it must be coupled with the capacity of workers to stop, or at least pause, work. That is to say, **workplace democracy includes the possibility of refusal**, and that possibility increases as we develop our critical consciousness. The possibility of refusal, thus, is a characteristic of workplace democracy that other researchers and practitioners can use in their interpretations of the workplace.

Of late, refusal has emerged as a vibrant theme in feminist data and data science approaches [329, 330, 145], with the Feminist Manifest-No as a prime example [331]. Refusal is also appearing in participatory design research and practice [332]. Data for Black Lives advocates for abolishing big data [333], which we might consider another form of refusal. In the context of DataWorks and similar programs, the disruption to the status quo brought about by critical consciousness and refusal will affect the program, the work, relationships, and the research. Accepting that disruption, and working within it, is necessary to sustain a commitment to a fair work environment that centers the workers.

What the workers at DataWorks engaged in is a tentative refusal, akin to a work stoppage. They did not categorically refuse to work on AAVE projects. Instead, the actions of the workers are those of the killjoy [334] in the context of research that Parvin and Pollack

[36] call out as necessary: to not give in to the aspirations of researchers or technology advocates, or accept so-called "unintended consequences" but to call attention to problematic situations and push back. In the face of perceived harm, the workers did not simply abide by the researchers' good intentions nor the possibility that these data sets and algorithms might be put to good use. Rather, the workers at DataWorks refused to continue working on the project until a series of conditions they authored were met. This turn of events is vital to understanding and fostering critical consciousness because it demonstrates the move from reflection to motivation and action: a move from awareness (which was achieved in the prior workshops) to developing a sense of agency to address oppressive conditions, and then working to change oppressive conditions.

From the outset, the goal of the workshops was to assist the workers in developing critical perspectives on data and algorithms in relation to their daily work at DataWorks. While the development of such perspectives might not be entirely surprising, the expression of this critical consciousness through halting a project and requesting the client make changes to a project to lessen harm was unexpected. What is important and distinctive is the move beyond awareness to action, from pointing out the significant problems with fundamental assumptions underlying the project to engaging in the act of refusal.

In demonstrating an arc and expression of critical consciousness, the AAVE project also brought to the surface the tensions that critical consciousness creates in the workplace. While some research has examined how critical consciousness and critical reflection may complicate the workplace [335, 336], less scholarship addresses how such complications might be productive for democracy. If we take democracy seriously (which we do), we must recognize and appreciate that a lack of tension does not characterize democracy. Indeed, while such tensions might be disruptive to the normal operations of the workplace, and therefore cast unfavorably by some, such tensions are essential to democracy. Following from perspectives of agonistic pluralism, democracy is characterized by the capacity for contestation [337, 338, 339]. Over the past two decades, theories of agonism have sig-

nificantly influenced Participatory Design; troubling research and practice by challenging assumptions about democracy and participation [340]. If the goal of democracy is not to arrive at consensus, then instead of designing tools or environments to hasten consensus, research suggests there is value in constructing events where contestation and dissensus can foment and be expressed. This is a different mode of participation than was previously standard to participatory design. Much of the work on agonism in participatory design focuses on community and government settings [341, 342, 343]. However, it could just as well be explored within the workplace, especially given the longstanding commitment of participatory design to work. Another characteristic of workplace democracy that researchers and practitioners can use in their interpretations of the workplace is agonism: a democratic workplace is one that allows for agonism. The question then becomes what tools and processes we might co-create to foster work environments where workers have the agency to challenge and reshape those environments?

Blending Context and Action

In discussing the issues of long-term and embedded participatory design in the Global Women's Fund, Trigg and Ishimaru [344] draw upon a set of distinctions first put forward by Gärtner and Wagner [345]. Those are the distinctions between Designing Work, Designing Organizational Frameworks for Action, and Designing the Industrial Relations Context. What is often meant by Designing Work is cooperatively designing the activities of work through cooperatively designing the tools used to perform the work. In the project described in this section, supporting workplace democracy occurred through an oscillation between the other two categories: Designing Organizational Frameworks for Action and Designing the Industrial Relations Context. This, in turn, connects us with the extension of participatory design into institutions and organizations as sites of design.

Designing Organizational Frameworks for Action "refers to the explicit and tacit norms on which organizational actors base their broader decisions" [345]. In DataWorks, these

norms included an emergent critical consciousness of data developed through the workshop. In other words, the organizational framework for action was—loosely—critical consciousness, and the design was in the workshop's structure, content, and enactment. At the start of the AAVE project, critical consciousness comprised awareness of the potential harms of data and algorithms and a sense of agency. But that awareness and agency had not been put into action because the workers at DataWorks had not yet encountered any substantially problematic jobs. It was only in the encounter with the AAVE project's offensive, biased, and harmful content that the move to action occurred, as the workers reacted to, and in particular against, that content. This reaction against is characteristic of agonism as described in the preceding section. This reaction against also surfaced a missing element of the workshops: a process for responding. Without such a process, the workers and management did what was most expedient and practical: to stop the work. The encounters and activities that followed, the meetings with the researchers, the authoring of requirements, and the request for a data use agreement blends into Designing the Industrial Relations Context. Through that process, the organizational framework for action that was loosely critical consciousness becomes much more fitted: the general qualities of critical consciousness are made specific to the context of data work at DataWorks.

Designing the Industrial Relations Context refers to "where legal and political frameworks are negotiated that govern the relations among industrial sectors" [345]. The process of working through the problematic conditions of the AAVE project and the eventual resolution of those conditions is an example of this negotiation work. In the AAVE project, the (industrial) relations are between the workers, the research staff and directors, and the clients, which in this case were other researchers. Those relations are power relations, making this an explicitly political endeavor [346]. What was negotiated in the AAVE project were the standards of working conditions. This included specific actions of data preparation taken by the client to remove offensive terms and phrases, an in-depth explanation of the work and the intended value of the work from the client to the workers, and the client

signing a data use agreement written by the workers to set boundaries for the use of the data. These requests and structures are essential because they assert respect and authority to the workers. Such respect and authority are usually absent from data work. Establishing these conditions as standards advances workplace democracy by acknowledging the workers' value and expertise. The process, more generally, demonstrates a shift in power relations between workers and clients. Of course, this shift is bounded and tenuous, but it is nonetheless a moment of achievement through contention.

At the same time, we should acknowledge that describing this process as *design* is a stretch. We intended the workshops to develop critical perspectives, but we did not plan for the AAVE project to unfold as it did. Indeed, as mentioned, we did not have a plan for responding to such problematic conditions within the context of a DataWorks project. As we consider this move into Designing the Industrial Relations Context through DataWorks we must acknowledge how we—as researchers—are responsible for and imbricated in this context and the challenges this raises due to contrasting positionalities.

Institutions and Positionalities

As an experimental workplace, DataWorks is a space to explore how else we might structure data work, and what sorts of institutions and organizations we might create to support more varied approaches to data science and computing. We—as researchers—are part of that exploration; we are part of those institutions and organizations. So, reflecting on how our work cultures and practices align and contrast with DataWorks and the workers is appropriate and important. Such reflexivity follows work in participatory design that examines the roles of design in navigating and influencing institutions [347, 348] as well as work in computing research calling for attending to the positionality of researchers [63, 349, 324].

One challenge we faced throughout the AAVE project was the instability that critical consciousness introduces into the workplace. The starting, stopping, restarting, stopping

again, and then restarting significantly disrupted the workflow in DataWorks. We had to reschedule client work and find other work to fill the days and weeks during the stoppages. In addition to this being an administrative difficulty, it also threatened to become a financial difficulty. Although there is a reserve operating budget we can draw upon so that we do not have to be on client work constantly, we do need to consistently complete contract work. As the work was delayed, we grew concerned, and the situation created a tension between the interests of the workers and the researchers. While that seemingly did not inhibit the expressions and actions of the workers, they were aware of the disruption, difficulty, and pressure this put upon DataWorks. Such problems are documented in the literature on critical reflection and critical consciousness in the workplace [335, 336]. Taking the cue to reflect critically upon our interpretation of the situation, we—the academics—were operating from a perspective that prioritized predictability. But if we are committed to democracy in the workplace, such managerial perspectives must be balanced with those that acknowledge the contestation inherent to democracy and we must continue to hold space for that contestation.

The process of stopping the work, bringing in the academic client, and the negotiations laid bare power dynamics within academic research specifically, and more generally, between those who request and those who perform data services. Outside of established formal institutions such as the Institutional Review Board, researchers are rarely directly questioned about the ethics of their work in conducting their research. Questions may be raised upon publication or other forms of dissemination, but rarely as the work is being done. Thus, the workers' resistance to the data work challenged not just the content and activity of the work, but also the traditions and routines of academic research and the power hierarchies within those traditions and routines. To the credit of the academic client, they met these challenges with openness: one reason they chose to employ DataWorks for the labeling was to work with data workers who would bring their positionality to the interpretation of the data. This returns us to an initial motivation for DataWorks: to engage

the university as a distinctive institutional site and in the process challenge, and attempt to shift, if ever so slightly, university cultures. In the context of higher education institutions, part of the work of workplace democracy is to push against the privilege of the university and the de facto authority given to academics, and instead work towards respecting the diversity of labor, skill, and expertise within and beyond the university.

While the data workers had the agency and motivation to act, they did not, *at the time*, have full authority or access in the university setting. For instance, before these projects, the workers interacted with clients but less regularly; most interactions with the clients were handled by a manager who was separate from the workers. So, in the events described in this section, the research faculty moved the emergent critical consciousness of the workers into the organizational structure and culture of the institution and served as initial facilitators with the client. In contemporary participatory design literature, this work is referred to as institutioning." [347, 348] Huybrechts, Benesch, and Geib describe institutioning as "a practice of interweaving between—as well as producing—various insides and outsides. . . consolidating and challenging existing institutional frames as well as forming new ones" [347]. In the case of the AAVE project, this institutioning involved bringing the workers and clients together, holding the space for contention, and assisting in establishing a process that foregrounded the workers in that project and going forward. This is not to valorize the work of the research staff or management—we made many mistakes. Rather, it calls attention to the necessary transformation of the role of the researcher in contexts such as DataWorks. Such institutioning requires more than being an ally: it requires becoming an accomplice [350].

Throughout the workshops and the AAVE project, the Whiteness of most of the research staff was a persistent tension. We attempted to ground the critical data literacy workshops in the lived experience of the data workers, who at the time were all Black. Similarly, we approached this work as what we perceived to be an opportunity for design justice: an opportunity to approach data work and the development of sophisticated data sets and

algorithms from the perspective of "nothing about us, without us" [326]. What we did not consider was how such experiences would be stressful, perhaps even traumatizing to the workers. **This was a failure of ours.** A few months after the AAVE project, the DataWorks manager left to take another job, and we replaced them by promoting one of the DataWorks workers, thereby putting a worker and person of color in a direct leadership role. As we iterate on the workshop materials, we are collaborating with the workers to select the sample data sets and examples that we draw upon, and future instantiations of the workshop will be co-taught with a worker. Acknowledging these problems and their consequences is crucial both to be more just in our research and also to further workplace democracy: we cannot solve the problems of Whiteness through projects such as DataWorks. Still, we can, and we must, hold ourselves accountable for how our work furthers these problems [351, 352, 353].

Forever Changing and Moving Forward

As an experimental workplace, DataWorks is forever changing. We—the research faculty and the workers—are constantly learning from each other and our collective experiences with DataWorks, adjusting and adapting how we work and what work we do. The experiences of the AAVE project have had significant effects on DataWorks, impacting our structure and processes.

The AAVE project provided an opportunity to reflect on the manager's role and consider how we might continue to transition more authority and power to the workers. We decided from then onward to promote a manager from among the workers. The first such manager was, in fact, one of the workers who had most strongly led in resisting the AAVE work. That manager left approximately six months later, after securing another position (a success for DataWorks). A new manager was then appointed from the current workers. This process of promoting data workers as managers puts the workers in direct control of their daily and weekly tasks, as well as putting a worker in the first line of decision making

regarding new clients and new projects.

We have also formalized the workshop curriculum, making it a component of the training process for all new data workers. As new data workers are brought in DataWorks and are learning basic skills in data wrangling, that learning is coupled with the workshop activities introducing workers to critical perspectives on data and machine learning, and uses the AAVE project as an example. That workshop curriculum is freely available online at https://dataworkforce.gatech.edu/data-tools/critical-data-literacy/. As part of formalizing the curriculum, we have also changed the examples and presentation of examples to address the trauma they might elicit, and we have provided the space within the workshop to better support the affective dimensions of critical consciousness through discussion and reflection.

We have been less successful in creating a standardized review process or binding and enforceable contracts. When a client approaches DataWorks, we request an explanation of their intended use of the data. This is presented by the manager (now a former data worker) to the other data workers. The workers, together with the manager, can raise concerns about a project and choose not to pursue a project. For instance, since the project described in this section, DataWorks has been presented with several other AAVE projects, which have been declined. Because each project is so different, checklist-style reviews are inadequate. Instead, we approach each project individually and assess it through discussion. Binding and enforceable contracts are also elusive. In large part, this is because there are scant data policies that would give substance to such a contract. Nonetheless, we do continue with data agreements, if for no other reason, because they are statements on behalf of the data workers at DataWorks of what they value as being the responsible use of data they collect and process.

4.2.3 Conclusion

Critical consciousness and workplace democracy can be nebulous concepts. This paper presented a case of the development of critical consciousness and workplace democracy within an experimental workplace. In doing so, we contributed ethnographic descriptions of these concepts in action and provided interpretations that connect to themes in participatory design and critical studies of data work. Specifically, we identified the role of agonism in the democratic workplace and described acts of refusal as part of workplace democracy, expanding the repertoire of agonistic data practices [354]. We also discussed the blending of designing organizational frameworks for action and designing the industrial relations context in data work. Finally, we addressed how multiple and varied positionalities affect this work. In future work, we plan to delve further into these positionalities and the institutional labor of establishing and sustaining workplaces, such as DataWorks, that strive to be just and fair within compromised contexts [355].

Of course, settings like DataWorks are different from most data work environments. DataWorks is an intentional exploration of how data work might be structured and performed otherwise, in ways that center the workers and their lived experiences. It is also set within a university, which is a particular institutional structure and culture. Nonetheless, the setting of DataWorks, and the lessons learned from such experimental workplaces are valuable to help us, collaboratively and collectively, understand what is possible, desirable, and undesirable. As researchers across the fields of human-centered computing, critical data studies, and design collaborate with workers in different ways to investigate and co-create more varied and equitable work environments, we hope that the descriptions, interpretations, and themes in this section contribute to those efforts.

4.3 Datum Fieldnotes: Enabling Dataset Documentation and Annotation in Spreadsheet-based Work

In this section, I introduce a novel Google Sheets add-on, *Datum Fieldnotes*, to automate documentation of per-datum dataset transformations and enable enhanced dataset annotation by data workers. Datum Fieldnotes is the core artifact and design product for a larger research-through-design (RtD) project [356]. The goal of this RtD project is to understand the experiences of professionals in the civic and non-profit (CNP) section whose jobs require extensive data work.

This section proceeds with a high-level motivation for the tool, followed by specific design implications based on workplace ethnography with DataWorks and existing challenges surfaced by related work. The first version of the tool is then presented, followed by a suite of user testing, the analysis of which informs the second version of the tool. The findings, or discussion of participants' experiences using the tool, are in keeping with the RtD style; that is, they are high level and focus on the prior experiences and perspectives of the individuals using them, as much as on the specific interface and affordances of the tool itself.

I then present three sets of next steps undertaken in response to user testing. First, a version 2.0 with improved user features. Second, a robust curriculum featuring everything from basic user guides to advanced data analysis modules that demonstrate *how* a dataset's *paradata* [57] can be used to establish and understand the dataset's own genesis. Third, the results of a focus group with DataWorks Data Fellows to further explore confusion and uncertainty raised by the terminology in the Datasheets for Datasets (D4D), as experienced by participants in the user testing. The results of this discussion are more fine-grained than than of the user testing described earlier; here, we are attuned to specific word swaps for terms introduced by the D4D project.

Finally, this section ends with a discussion of design implications for a version 3.0 of

Datum Fieldnotes' interface, as well as continued work on making D4D tractable for data workers, with the goal of having data workers participate in the completion of datasheets for datasets, as part of the early auditing process.

4.3.1 Why integrate perspective into data processing?

The prior sections of this chapter explore the role of data workers' lived experience and perspective in shaping datasets. Critically, from the case study above (4.1), one element of being able to contest the means of production of a dataset—or for a data worker to interrogate their own role in producing that dataset—is having the evidence to back up their position, in the form of examples in the dataset, or features to point to. In reflecting on this experience, as well as the needs of civic and non-governmental data workers (as operating in professional environments with limited resources and needs for collaborative data work environments, see 3.2), I saw a need for tools that expressly supported flagging questions or curiosities in a dataset. How can we make it easier for data workers to catalog and archive their observations, reflections, and questions about a given dataset?

These dataset annotations, provided by data workers are paramount to complicating our notion of what a given dataset represents. As Bowker reminds us, "data is never raw," [273] but when it arrives in a spreadsheet, not yet cleaned or standardized, it can give the appearance of an unpasteurized egg purchased from a commercial supermarket; on the surface, its aura of distinct origin (feathers, bits of debris clinging to its shell) can mask the complex sociotechnical process by which that given egg arrived on the grocery store shelf. There are the vast—potentially even global–transportation networks and trade agreements between stores and farms. There is the agricultural shift from subsistence to commercial farming. There are the cultural and chemical arguments for and against cleaning or sanitizing the egg's exterior. Each step of this process uniquely facilitated this egg's appearance on the grocery store shelf.

The unpasteurized egg metaphor serves as a means of explanation for why we need

tools to allow us to interrogate a dataset in situ, or as it currently exists to us. Here, however, the debris takes the form of dataset observations and feedback from data workers, as well as *paradata* (or process data) about the development of the dataset itself [58]. Further, when that dataset is shared between, say, co-workers or organizational peers, those interrogations—as annotations—should be included. Specifically, they should be attached the very dataset itself. This is the motivating rationale behind the tool I introduce in this chapter, or *Datum Fieldnotes,* a Google Sheets add-on that allows users to engage with a dataset's metadocumentation as a dataset itself. Datum Fieldnotes is a tool to help us understand a dataset in context and to depict data workers as *designers* of a dataset. In depicting the data workers as a designer, we are consistent with Feinberg's notion of data as *designed,* or intentionally constructed archive of a sociotechnical process, one necessarily shaped by the perspective and tools of the people creating it [232].

In more concrete terms, Datum Fieldnotes does two things for users. First, it records changes on the per-datum level, or per change to any given cell in the spreadsheet. This is unlike most conventional spreadsheet software documentation tools, including Google Sheet's own tracking feature ("Version History") which saves batches of changes to a spreadsheet, rather than individual ones. Further, Datum Fieldnotes explains the change that took place, describing the type of change (e.g., manual or records a function, if used) so that the user can understand what kind of data transformation took place. Second, it records those data transformations into an accessible, queryable log that encourages engaging with a dataset's development information as a dataset itself. This can be used to help collect and archive a dataset's originating context. In particular, it can support efforts like the datasheets for datasets project [288]. As the questions on the datasheets call for information such as who was involved in the data annotation and creation of the dataset, as well as more nuanced questions about the subject of the dataset, the automatic documentation of the Datum Fieldnotes supports a user in answering.

Turning specifically to civic and non-profit (CNP) organizations, we encounter a unique

set of data work challenges. CNP work is mission critical, as it is needed both to track and demonstrate organizational efficiency, within the organization and to funders [357], and yet it can take away time and resources from enacting the key mission of the organization, such as direct contact with clients [50]. Unlike data work that happens in for-profit organizations, CNP data workplaces are often characterized by a lean budget and constant workforce fluctuations, as a result of volunteer labor and donation-dependent staffing capacity [358]. At the same time, the data work that is done in CNP organizations is high-stakes; the ethical concerns related to doing related work properly (e.g., avoiding missing or mis-inputting data) stem from the organizational mission, which is directly concerned with improved the experience of daily life from clients from various populations in need, such as fighting homelessness [359]. At the same time, many commercial-grade data work platforms, such as CRMs [360], often require both expensive license fees and dedicated staff members to learn and/or maintain them [361, 362].

Traditional spreadsheet software has been identified in past literature as a more sustainable alternative, given the low barriers to entry and low or no cost license fees [363]. Spreadsheets, however, lack the advanced dataset documentation features needed to make actionable *paradata*, or process information about the development of a dataset that can be used to contextualize a dataset and determine appropriate future use for analysis or system development [58]. Paradata as a tool towards ethical data practice is especially important for CNP data, which is generally focused on adding value to human life or experience. Consider an organization serving clients who experience homelessness; missing or uncertain data can impact material resources allocated to an individual client or the organization as a whole.

Best practices for working with data require first knowing how to understand and interpret the dataset a user is working with, which has increasingly become a core competency for democratic participation [364]. Previous research in HCI has shown that CNP data workers already have robust dataset contextualization practices [365, 50, 359] that can re-

sult in workable paradata, if properly archived and shared. What is missing is the tools to streamline and expedite this process, without adding to the heavy data work burden these data professionals already undertake. Our Google Sheets add-on, Datum Fieldnotes, is designed to help capture and formalize those dataset contextualization practices. Datum Fieldnotes has two key functionalities. First, it automates micro-documentation of user changes per datum (or, cell-level) in the spreadsheet, unlike the in-built features of conventional spreadsheet software. Second, it provides an interface for users to create long-form notes, or records of questions, concerns, or observations about an individual dataset value or datum transformation. It is the creation of this record of dataset creation and archiving of reflections and observations that foregrounds the process by which a dataset comes into being. This record of genesis, in turn, enables interrogation and contextualization of the dataset itself, which are needed to data-intensive systems that will behave as expected [366, 367, 368].

Datum Fieldntoes was developed as an artifact for a research-through-design process, as part of a bigger investigation focused on the development of tools, systems, and workflows for the development of better-contextualized data work. In this late breaking work, we focus on the design and testing of the tool itself, as a case study in the development of a data work tool specifically geared towards CNP workers, as an oft-neglected data work population. We share the initial design of the tool, the structure and results of a user testing study with CNP workers, and resultant design changes and supplemental learning materials developed. We focus on the tool as a means of facilitating *data infrastructure literacy* and how design choices can support deeper dataset interrogation, bringing the background work of data work to the forefront [368]. This is a crucial support in the world of non-profit service organizations, where funding requirements and collective impacts are driving factors for efficiency and effectiveness, engaging them in "data confidence" and "data compliance" work [369].

In performing this work, we had two research questions. **RQ1**: can we offload some of

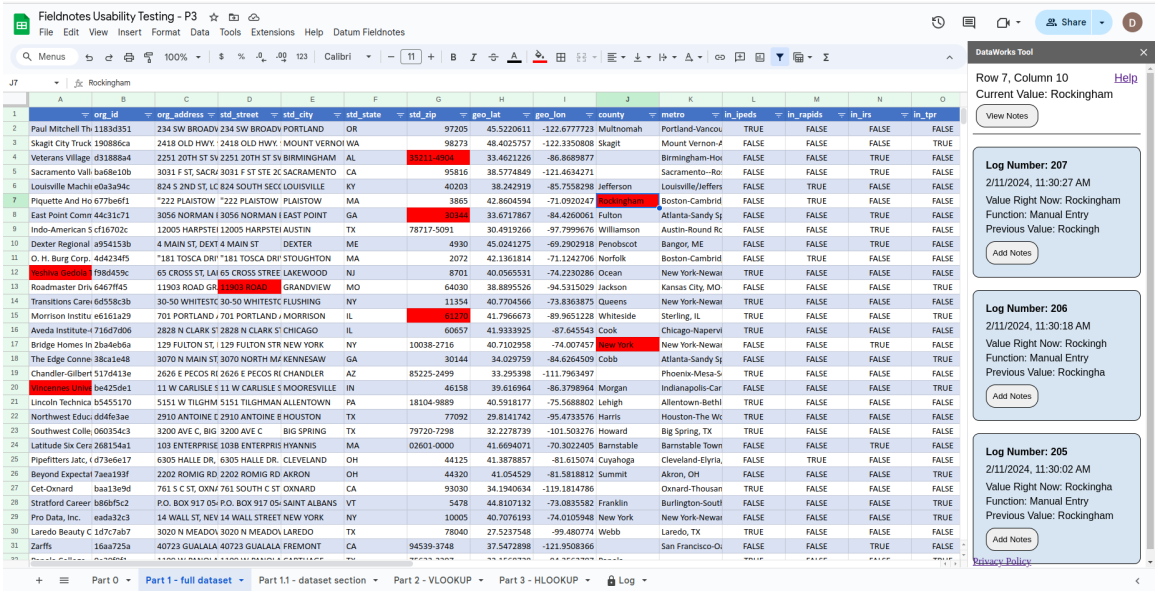


Figure 4.1: Image of the add-on, once loaded, showing the recent edits to a given cell in the current spreadsheet.

the additional labor needed to document and archive the paradata that CNP data workers produce? And, subsequently, **RQ2**: how can we facilitate CNP data conextualization in action, to make sense of what other system features can help concretize this work? Our work offers addresses the urgent call to develop datasets that can build trust, reduce antagonism, and help hold a mirror to organizational or governmental values, to serve as a potential site of data contestation [366]. These practices are critical to developing responsible data-intensive systems, or ones produced off of datasets that are both well-understood and properly contextualized [145, 370]. Further, Datum Fieldnotes provides methods to centralize and record dataset authorship [371, 372], pointing to data workers as domain experts, who are best positioned to determine fair and appropriate use, in an era of rife data misuse [373].

4.3.2 Datum Fieldnotes version 1.0

The CNP data work design space

This work draws on two related areas: first, the data work burdens and challenges that civic and non-profit groups (CNPs) undergo. Second, the utility of research-through-design (RTD) as a method for studying data work in CNP organizations.

The CNP Data Burden. CNPs rely on data work to demonstrate both organizational impact and potential to external funding bodies [374] but also to track progress towards internal goals (or benchmarks) [357] and identify relevant mission gaps or opportunities through open data [361]. However, there are several barriers for adoption of new data work tools or systems for CNPs [375]. There is high staff and volunteer turnover in many cases—due to grant-dependent positions and ad-hoc volunteering, respectively—which poses challenges for the maintenance of internal knowledge of data systems [362, 374]. When CNPs do have capacity for new data systems adoption, those systems are often developed by volunteers or staff members who are experts in the mission domain, rather than data systems technology, resulting in what Voida et al. term homebrew databases [374]. Designed data manger staff members and advanced data work system and tool subscriptions count against the administrative budget or "overhead" on which non-profits, in particular, are graded. Those grades are used by potential donors determine the organization's fiduciary prowess [376].

The emphasis on CNPs' data can also re-direct organizational mission as a result of "data drift," where data findings push the mission in a new directions [267]. This creates an endless cycle of data work that indirectly supports the organization's mission (by helping accrue funding) [377], which subsequently requires more program evaluation data work [359]. Further, the burden of data work can also get in the way of the carrying out of the work directly related to the CNPs mission [50], for example, performing advocacy work.

In short, improved systems must be flexible to the organizational mission, but require

little customization, as the more nuanced or advanced a potential software solution is, the less likely it is to be usable in the long term. There is also motivation to not update or upgrade software, or improve otherwise improve data systems, that would incur expense, either in subscriptions or staff time, since this counts against CNPs overhead costs. This leaves CNPs in the lurch, resulting in data systems that are often "fragmentation and duplication" of data systems that render them useless in the long-term, or even result in incorrect data analysis [369].

Research Through Design for Studying CNP Data Work. Given the pressures on CNPs to produce data for funding bodies and for exchange with other related organizations, data work becomes a major task, but one that is a difficult task to tackle [378], in part, because little attention has been paid to CNP data workers, relative to their counterparts in commercial data science [365].

While these concerns have been documented by related work, there is a need to focus on applied methods for improving data practices in CNPs, both to support CNPs in getting the funding that they need [377] and help them determine program efficacy and leverage existing open data resources [361]. We chose to employ the RTD method [379] because, to elicit the minutiae and nuance of CNP data work, we needed to see it happening in real time. Our goal is not only to understand CNP data work practices, however our goal is to create something useful for the CNPs that we work with, as part of a larger research alliance. Engaging in this infrastructural inversion, or bringing the background work of CNP data processing to the forefront [161], allows us to reflect on the social implications of CNP data work production and what the effect funding bodies' emphasis on quantification has on CNPs more broadly.

In particular, we wanted to know how we could support the data infrastructure literacies of our partner CNPs, or whether our partners needed support in understanding how a data system might align with the motivations for system design or dataset creation and use [357]. This is a goal of our larger research project, in which we are focused on supporting CNPs

in making best use of their data and contextualizing the development and data stories that emerge from their datasets [361]. We employ a critical technical practice approach [380], aiming to assist CNPs in auditing their creation and use of their own datasets. Given that many CNP datasets end up as open datasets, these are increasingly subject to use by AI and ML system developers. Our goal is to enable CNPs to exert ownership over the data narratives that emerge and ensure that any predictive or generative system trained off these datasets matches the organizational mission of the CNPs that produced it.

Tool features and structure

The first version of Datum Fieldnotes is based on the known needs of data workers in CNP organizations, based on our research team's multi-year workplace ethnography with a data services provider, DataWorks, as well as known concerns raised by studies of CNP data work (see 4.3.2). DataWorks provides contract labor for several CNP clients, focusing on early-stage data work. Drawing on this knowledge, we identified five critical, co-located implications for design resulting in the overall shape of the Datum Fieldnotes tool. The tool is 1) built on top of spreadsheet software, that is, 2) affordable and 3) easily accessible without vast technical background. It must allow for 4) dataset documentation to travel with the dataset itself and, 5) follow industry-grade data documentation practices.

Datum Fieldnotes: Spreadsheet Add-On for Google Sheets. CNP workers need data work tools that are accessible, adaptable, and replicable [363]. While Excel is commonly used for many businesses and organizations, we know that many CNP groups utilize Google Sheets, in part due to the collaborative-first nature of the service, and the lack of license fees (and time requirement for applying for license-fee waivers) [381]. Further, Google Sheets closely aligns with Shaprio et al.'s design tenets [363]. As Sheets is a cloud-first spreadsheet service, integrated with the existing Google Workplace Suite—already used by many CNPs—sharing access and security is handled by Google, rather than individual organizations [362]. Sheets already has a plethora of online resources and support and was

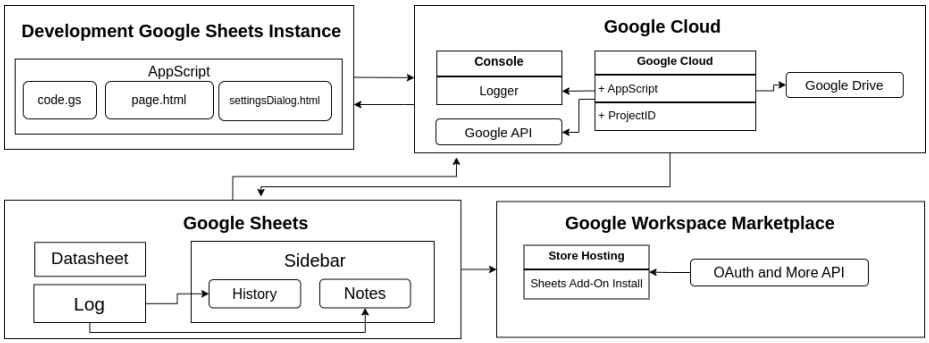


Figure 4.2: Information flow diagram for the Datum Fieldnotes tool.

already used by many CNP clients of DataWorks. For those CNP data workers unfamiliar with Google Sheets, the existing learning materials minimizes training overhead [359, 382].

Sidebar: Approachable functionality for data workers' observations, reflections, and questions. In the design of Datum Fieldnotes, we prioritize user comfort with documentation input and attempt to automate the process of micro-documentation, rather than asking CNP workers to double their task load. The key input feature is the Notes functionality which occurs in the Sidebar, while the Log displays both those manual annotations, as well as a record of all individual data transactions. The Sidebar opens when a user opens the tool on a new sheet and is populated with up to five changes from that session (or since the user last opened the Sheets project). The recent displayed correspond to the cell that the user is currently interacting with, as the user performs a data transformation on any cell in the dataset. The Sidebar is a graphical user interface that sends information to the Log, enabling users to see and interact with information about recent spreadsheet changes. Users can input notes, or observations, comments, or questions they may have about a cell's value, via the "Notes" feature, housed on the Sidebar. Notes allow users to create textual reflections on a particular cell or individual data transaction (cell change), contextualizing a value, or flagging a particular issue within the dataset. Once a note is created, the cell in which the note lives is highlighted for future users and for a data worker's own later reference, and is recorded to the Log.

Log: Automating micro-documentation in the background. Upon first opening the

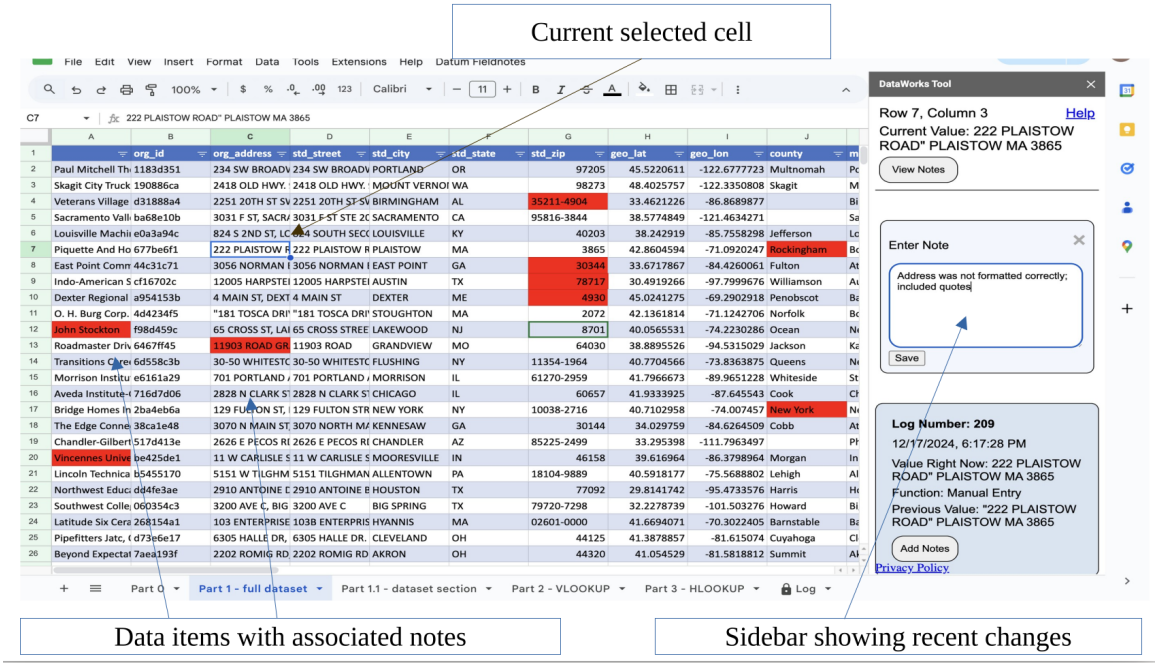


Figure 4.3: Datum Fieldnotes 1.0 interface with annotated features.

tool, a separate tab in the sheet (workbook) is created labeled as the "Log". For every edit made to a cell in the sheet in which the tool is active (including adding a Note), the Log will have an entry added with essential information about the change, including where the change was made, what the change was, when it was made, and who made it [377, 383]. This information is stored chronologically in the Log, which can be easily queried for a particular change, author, or timestamp, automating the collection of interrogatable *paradata* [58]. When data is missing or incorrectly reflects events or individuals, these discrepancies become ethical challenges in fair dataset use and distribution. This propagation of ethical shortcomings is particularly concerning in the CNP context, where the focus is on understanding and allocating resources to populations in need [359]. Our Log provides one method to address these issues by providing a central location to double-check work done on a dataset and review all dataset alterations (or, individual data transactions) made. This change-by-change version history adds more specificity and detail than the standard version history that Google sheets keeps a record of automatically.

Log: Workable paradata, exportable as CSV. As with any Sheets project, when

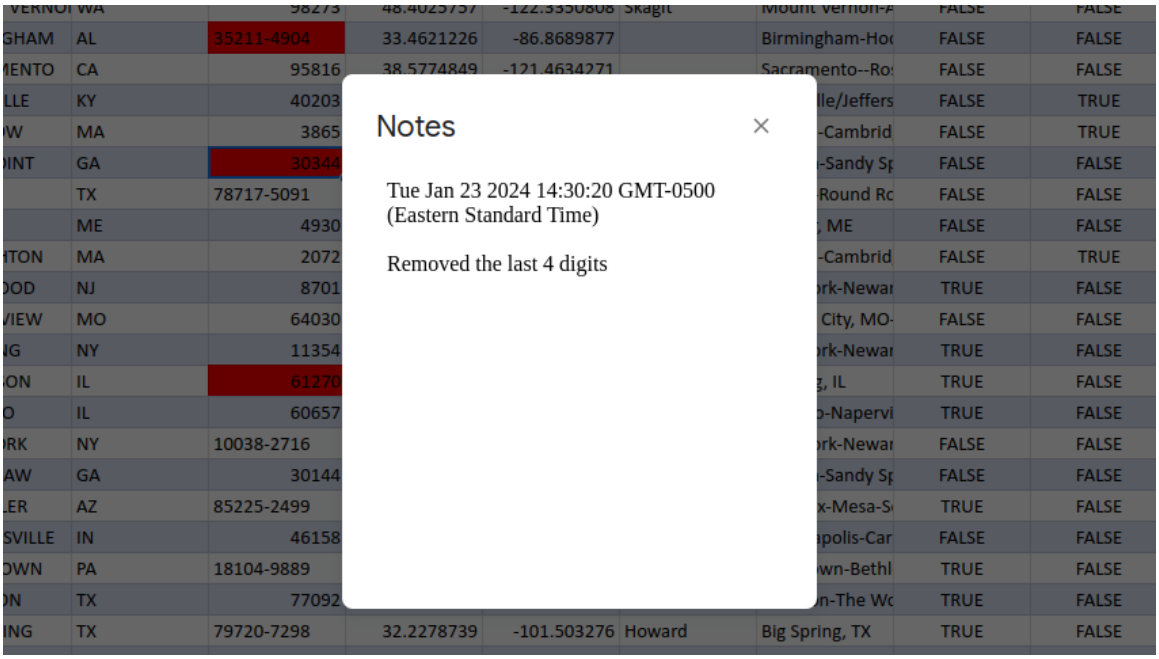
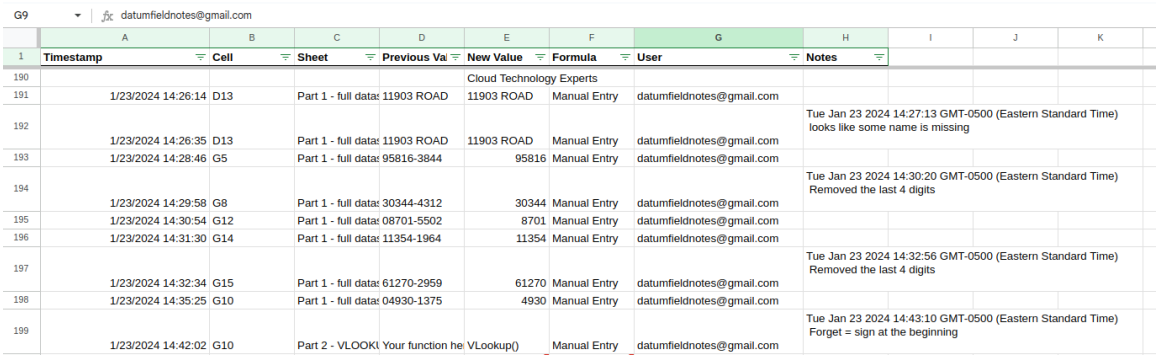


Figure 4.4: Example of what queuing up prior notes on a cell looks like in Datum Fieldnotes 1.0.



	Timestamp	Cell	Sheet	Previous Val	New Value	Formula	User	Notes
190					Cloud Technology Experts			
191	1/23/2024 14:26:14	D13	Part 1 - full data	11903 ROAD	11903 ROAD	Manual Entry	datumfieldnotes@gmail.com	
192	1/23/2024 14:26:35	D13	Part 1 - full data	11903 ROAD	11903 ROAD	Manual Entry	datumfieldnotes@gmail.com	Tue Jan 23 2024 14:27:13 GMT-0500 (Eastern Standard Time) looks like some name is missing
193	1/23/2024 14:28:46	G5	Part 1 - full data	95816-3844	95816	Manual Entry	datumfieldnotes@gmail.com	
194	1/23/2024 14:29:58	G8	Part 1 - full data	30344-4312	30344	Manual Entry	datumfieldnotes@gmail.com	Tue Jan 23 2024 14:30:20 GMT-0500 (Eastern Standard Time) Removed the last 4 digits
195	1/23/2024 14:30:54	G12	Part 1 - full data	08701-5502	8701	Manual Entry	datumfieldnotes@gmail.com	
196	1/23/2024 14:31:30	G14	Part 1 - full data	11354-1964	11354	Manual Entry	datumfieldnotes@gmail.com	
197	1/23/2024 14:32:34	G15	Part 1 - full data	61270-2959	61270	Manual Entry	datumfieldnotes@gmail.com	Tue Jan 23 2024 14:32:56 GMT-0500 (Eastern Standard Time) Removed the last 4 digits
198	1/23/2024 14:35:25	G10	Part 1 - full data	04930-1375	4930	Manual Entry	datumfieldnotes@gmail.com	
199	1/23/2024 14:42:02	G10	Part 2 - VLOOKU	Your function he	VLookup()	Manual Entry	datumfieldnotes@gmail.com	Tue Jan 23 2024 14:43:10 GMT-0500 (Eastern Standard Time) Forget = sign at the beginning

Figure 4.5: Excerpt from the the Datum Fieldnotes 1.0 log that shows all cell-level data transformations to the spreadsheet.

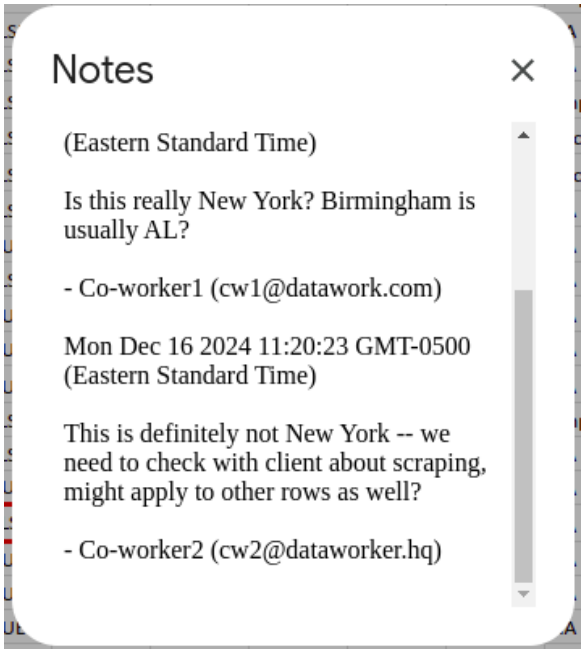


Figure 4.6: Datum Fieldnotes 1.0 notes modal.

the tab on which the tool is active is shared, the Log sheet is shared along with it. The Log is also a fully exportable as a CSV, meaning the dataset paradata can travel natively with the dataset itself, unlike a separate README document or datasheet. This enables cross-platform dataset sharing and utilization [384]. Non-profit specific data work is characterized by a need for flexibility and is an important consequence of the influencing factors of funders, community, and environment [202, 50, 361]. The ability to apply work done in the Datum Fieldnotes tool to other systems enables wider use of the added adaptability, as, for example, a data analyst working with a housing non-profit could share their dataset (either as a Google Sheet or CSV) with city officials, contractors, or other actors in the same field without having to worry about compatibility issues or missing information, with the paradata contained in the Log seamlessly accompanying the main dataset itself.

Datasheets for Datasets integration: Cross-industry and downstream-aware documentation. There is an increasing movement to summarize the a dataset's origin and intended use in the form of a datasheet, inspired by the documents that accompany hardware fixtures to explain their capacities and capabilities [288]. These datasheets for datasets

(D4D) have been shown to support, for example, ML engineers notice and engage with the ethical issues in training data for ML systems [385]. In this first version of the tool, we included a few hand-selected questions from D4D relating to the development and intended use of the dataset, which autopopulate in a new tab of Sheets project when a user opens the Fieldnotes extension. The incorporation of D4D questions promotes further contextualization and reflection by data workers—who are experts in a dataset's contents and formation—to serve as guides for later use in potential development of generative or predictive systems trained off a dataset.

4.3.3 User studies and in situ testing

Methods

To evaluate our Datum Fieldnotes in context, we conducted twelve IRB-approved usability tests with the initial version of the tool, with participants who work with CNP data.

We recruited interviewees from both DataWorks and our own network of CNP community partners. When participants signed up to participate, we asked them to give a description of the kind of spreadsheet datasets they usual work with. We then use ChatGPT to create a synthetic dataset that emulated the genre of data (e.g., ecological measurements collected from sensors), so that participants were not required to share their own organizational data. The research team annotated the dataset as if we were their colleagues, adding questions, comments, and changes, that, in the second part of the user testing session, a contextual inquiry activity [386], in which, after a short demo of the prototype, we asked participants to trace, to make sense of their imaginary colleague's concerns, reflections, or motivations in the data work they had conducted. The first part of the session included a brief semi-structured interview in which participants were asked about their background with spreadsheet-based data work.

The first seven participants were from DataWorks (referred to with the prefix 'DW #'). These participants were not directly compensated; the study was voluntary and participants

were indirectly compensated with their standard hourly wage (research engagement is a compensated part of their roles). The other five participants are CNP practitioners (referenced as 'P#'), working in fields from support for veterans to eradicating hunger. These participants were compensated with $25 Amazon gift cards. All participants identified data work—specifically spreadsheet work—as a core competency of their work, though spreadsheet were used to collect and analyze everything from numerical distribution of material resources for clients to thematic analysis of transcribed oral histories from community residents.

Testing took place via both in-person sessions and virtually via a recorded Zoom call. In all cases but one (resulting from machine incompatibility), the researcher's desktop was used, to avoid unintentional exposure of participants' personal or work data. The research team then engaged in thematic analysis [387] of the recorded sessions, beginning with transcribing the sessions and re-watching all interviews. The first, second, and third authors engaged in memo-ing, pooling their notes collected during sessions, post-session reflections, and re-watching session recordings and created a preliminary codebook, which was then refined through subsequent rounds of analysis and discussion. This resulted in a final codebook that was applied to all transcripts.

Participant Experiences and Reflections

Given the RtD impetus of this project and subsequent tool development and user testing, the results described here focus on participants' descriptions of experiences and perspective as elicited by the experience of using the Datum Fieldnotes tool. Further, encounters with the tool encouraged them to recollect and share experiences of data work and obstacles experienced therein that may not have come up in a more typical user interview.

At a high level, given all participants' prior experience with Google Sheets, participants reported that the combination of a familiar environment with the tool's simplistic design made them feel comfortable using the tool after a short walkthrough. When prompted to

complete discovery activities such as determining who had made a change, in all but a few cases it was the first instinct of the participant to go straight to the Log, suggesting that it was understood by participants as a locus information about the dataset. For example, DW 5 described the tool as being, *"...actually self-explanatory; it's very easy to use, which is good."* In the rest of this section, we describe the more nuanced experiences and reflections of participants.

- Documentation practices status quo. The problems participants had were consistent with what we hypothesized they would be, when designing the tool, particularly regarding data documentation and version control. These led to documentation disorganization (P5), repeating manual work (DW 5, DW 6, DW 3, DW 2), spending time understanding documentation of others, and translation getting lost (P2). Documentation practices varied between data workers. Some participants used external READMEs (e.g., separate Word documents) or other documents external to the spreadsheet environment in order to record high level changes to a dataset (DW 1, DW 2, DW 6). Difference within use of READMEs varied as well; sometimes they were full records of dataset changes and other times they were merely highlights to be shared with a client (DW 6). Others relied on keeping backup copies of the spreadsheet (e.g., P2), while yet others (e.g., P6) regularly changed their documentation methods depending on the needs of individual clients.

- Enabling enhanced communication, upsetting documentation color schematics. The additional "Notes" feature of the tool provided useful communication avenues for participants. Users noted how having records of changes and notes, with P3 saying it *"would have saved so much time"* when collaborating with others on a dataset. A few participants (DW P2, DW P4, P6) compared our notes to the built in Google Sheets comments, showing initial confusion over the benefit of their additional inclusion (suggesting a need for us to better differentiate the two), but came to find throughout our interview activities how these two methods of added documentation differ and

can be used in conjunction. Several (DW 1, DW 3, DW 4, DW 5, DW 6, P2, P3) participants described struggling with version control in collaborative spreadsheet-based work. Participants relied on ad-hoc methods of color coding, notes or comments features native to their spreadsheet software, and some participants maintained external central reference points. In particular, the color used to highlight a cell indicated a concern type, e.g., question, data type, or responsible team member. Both the native Google Sheets "comment" feature and Datum Fieldnote's highlight thus lack the required custom color assignment flexibility.

- Self-reflection, growth, and learning in the workplace. Several participants remarked on the Log and Notes' potential as a running "note-to-self" that would be helpful in remembering and keeping track of not just the transformations being enacted on the dataset, but how and why these changes were being made. This matters for CNP data workers who may be learning on the job (particularly volunteer data workers), with P1 explaining, "That's how we learn." Participants discovered functions incrementally as well as some users wanted to know to view notes. The DataWorks participant found this filter recreation task particularly applicable to their experience with more collaborative work and on-job learning, making note of how much easier it was to reference the log to find and copy a filter or function they needed as opposed to spending excess time figuring it out from scratch. Multiple participants were unfamiliar with advanced formulas-—or their formatting in Google Sheets, though the underlying concepts may not be new–and identified ways that the structure of the tool could better support them. For example, P2 struggled to use filters, as applied by their imaginary co-worker, because the logging of filters did not include the "=" sign that prefaces filtering operations. DW 4 echoed this, describing the struggle of learning a new operation during work, but then not having repeated opportunity to regularly apply it, and subsequently forgetting it. DW 3 also expressed confusion in the value of version control, particularly as enacted by the tool, suggesting the

need for more support of tool users learning new data concepts, not just new data transformation functions.

Participants' D4D Reflections

Another key finding was participants' experience completing the selected D4D questions. We asked them to theorize about the following ten questions, based on what they knew from interacting with the fabricated dataset. These questions were selected from the larger D4D set because we thought they could reasonably be approximated by participants, given that they focused, first, on the data work portion of AI system development, and, second, specifically interrogated the process of dataset development, which Datum Fieldnotes is designed to address.

- For what purpose was the dataset created?
- Who created the dataset?
- Is any information missing from individual instances?
- Are there any errors, sources of noise, or redundancies in the dataset?
- How was the data associated with each instance acquired?
- Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled/ that might impact future users?
- What data does each instance consist of?
- Does the dataset contain data that may be harmful or confidential?
- Are the individuals unintentionally identifiable by the data?
- Does the dates contain data that might be considered sensitive?

What we observe is that the terminology that emerges, when our participants describe their data work, is much more tied not only to data as a *tangible, 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 D4D 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 to 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 these 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.

While these changes may be overlooked or deemed insignificant by AI model designers and engineers, they are fundamental changes to the dataset that will shape the way the

contents and concepts that emerge from the dataset can be understood. Standardization is the difference between, for example, a series of trailing digits and a rounded number or count. If this rounding is not explicitly stated and considered, subsequent analysis could be inherently faulty. Rounding errors, while they may seem inconsequential in the vast majority of application settings, are not without potentially disastrous implications, such as the rounding error partially responsible for the implosion of the European Space Agency's Ariane 5 Rocket flight V88 [388].

In short, 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.

4.3.4 Acting on insights from user testing

Reflecting on the results of our user testing sessions, we found three actions were necessary. First, there exists a set of discrete technical changes necessary to make the tool more usable for CNP participants. Second, we needed to design and produce curricula that would support (and remind) CNP data workers in the tool's features and enable them to use the paradata produced in meaningful ways–in other words, to scaffold analyzing the paradata in ways that contributed to dataset contextualization more broadly. Third, we recognized a need to better understand the vernacular data work terms that emerged in the D4D portion of the user testing activities. To address this third point, we ran a follow up focus group with Data Fellows from DataWorks, with whom we discussed what more appropriate terminology and question phrasing would be, with two contributions. First, we share a focus group guide that can be adapted to individual organizations, to highlight the terminology and contextual knowledge that is most appropriate when discussing dataset concepts and changes. This evolves the D4D question set, to make it easier for data workers to engage with, so that data workers can, in turn, be engaged with as partners in the responsible AI pipeline and as dataset auditors. The second contribution is a set of findings that introduce

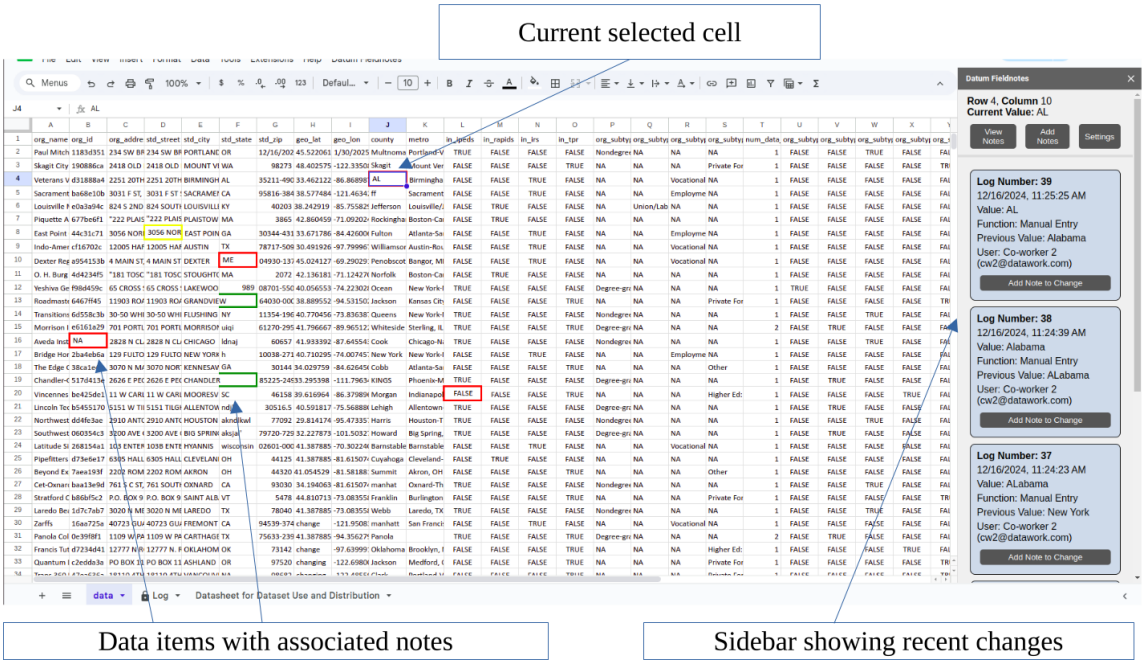


Figure 4.7: Datum Fieldnotes version 2.0 interface features.

sites of contextual differences in the genesis and development of datasets at different data work sites–for example, in the definition of relatively sensitive subject matter inclusion–presenting next steps for the design of data work tools that support the localized expertise of data workers.

Datum Fieldnotes Version 2.0

We made two significant changes to the design and interface of Datum Fieldnotes, along with assorted minor tweaks and bug fixes, that comprise the second version of the tool. First, we enhanced the data transparency and user-centric design by addressing users concern for accessible viewing and use of tool, through the introduction of new options for attributing changes by a given user in the Log. Second, we added further logging and documentation features and additional protection of user privacy. Finally, we added the complete D4D list of questions, as well as preliminary additional questions constructed to be familiar and completable by CNP data workers.

User-Centric Design and Accessibility. From our interviews we noticed that users

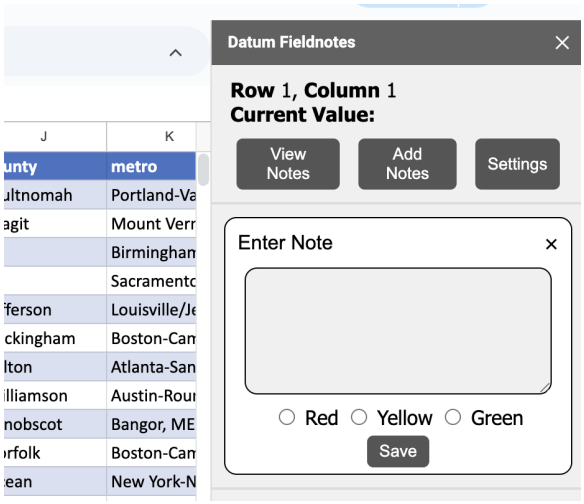


Figure 4.8: Datum Fieldnotes 2.0 sidebar interface, showing color styling options for cell highlighting.

needed more features to complete the documentation of cell changes, as well as in the tracking of changes to notes. This includes a need for custom color styling of annotated cells, since participants frequently noted that they used colors to denote different kinds of observations or changes, or to depict the current status of a cell. These changes are consistent with constant fight against dataset "fragmentation and duplication" [374], wherein rather than having to separate out parts of the dataset into different spreadsheets, so that multiple data workers can work synchronously, all changes can be made to the original spreadsheet directly. Further, to accommodate the cell highlighting schema used by participants in their own documentation methods, we switched from a full cell highlight to a colored cell border, which makes it possible for cell highlighting to still take place.

Further, the user interface has been redesigned in a palette of black, gray, and white, enhancing readability. In addition to feedback from participants, this also better aligns with the importance of accessibility and adaptability necessary to make CNP data tools useful [363].

Enhancement of Data Transparency. In this version of the tool, we introduced a customizable logging feature that allows users to choose how their contributions are attributed (name or email). This feature will allow data workers in privacy conscious settings like

schools [P4] and in government work [DW7] to balance transparency with autonomy. This design choice reflects the concern raised by Harmon et al. [362] regarding the sensitivity of data held by CNP organizations and the need for secure data management.

The new settings modal allows users to choose their preferred methods of logging their identity, through their name and/or email. If the user chooses email then the user is prompted to only enter their emails. This flexibility ensures data workers, like those at non-profit organizations, can maintain privacy while still ensuring there is sufficient information for follow-up, if necessary; we believe this facilitates a balance between data transparency and user autonomy.

We now also provide a session history that shows all changes made to a cell directly accessible through the side bar. A user can find information about a cell through the sidebar and 'Log' tab. Reflecting on the challenges of "homebrew" databases [374] like the ones our participants use, we designed this cell and change tracking to help data work collaborators trace who made what changes and how they did so, as well as provide potential contact information. This change supports data-driven collaboration for collaborators with varying levels of data literacy [359, 267], as is frequently the case in non-profit organizations. This change allows collaborators to trace the evolution of data, identify contributors and understand the context behind modifications. This level of accessible and transparency supports individuals with varying levels of data literacy [389, 368, 377].

Our emphasis on collaboration also extends to the way collaborators have their spreadsheet work and experiences recorded as part of the metadata. To further enhance context and promote responsible data reuse [383], we upgraded the D4D integration with the full set of questions from the paper [288], as well as a handful of questions that were mentioned by our participants that is broken into two main parts of table of contents, "For Data Workers" and "Datasheets for Datasets". The table of contents is linked to those sections of the Datasheet for easier access to parts of the Datasheet. The "For Data Workers" section addresses those who created and worked on the dataset, difficulties and patterns encountered

as well as smaller notes for how the dataset should not be used. The users who work on the dataset question are automatically updated on the datasheet logging users who make changes to the dataset. The section with the full list of D4D questions is a deeper dive into the context of the datasheets purpose, distribution and collection and processes that include working on the dataset. Participants like P4 and P2 noted that they have encountered datasets where they would've found dataset questions useful to finding context in the dataset they're working with. We updated the D4D updated with the full set of questions from the paper [288], as well as a handful of questions that were mentioned by our participants, into the sheet for seamless updates, facilitating structured metadata documentation and promoting data reuse [383].

Datum Fieldnotes Curricula

In response to the user testing, we developed a second version of the tool. These changes include 1) enhancement of data transparency and user-centric design, 2) further logging and documentation features, 3) support for installing and learning to use the tool (and resultant paradata). However, we focus on the the last of these, and its importance for making tools that are sustainable for CNP workers. The overarching finding was that, yes, our tool was usable for CNP data workers after a short instruction, but to make the next step, we needed to be better support them in continuing to develop their rich contextualization analysis practices. Namely, throughout our analysis, we found that the benefits offered by the tool could only be enacted when participants had ready access to robust documentation and refresher material. In the CNP context, we argue creating these materials is actually the more important follow-up contribution, given the fluctuation in data worker population and skills, than technical improvements on the tool itself.

As a result, we developed an extensive online curriculum [1], ranging from simple guides on how to install Datum Fieldnotes and employ its core features, to a series of videos

[1] https://dataworkforce.gatech.edu/datum-fieldnotes/

on how to perform analysis on the paradata recorded to the log. The curriculum starts with simple guides on how to how to install Datum Fieldnotes. It then goes on to explain the core feature guides and how to use them which addresses the fact that many users expressed a lack of familiarity with data tools and the specific functionalities of the tool. Users repeatedly expressed the need for clear explanation on how the tool functions [DW 4, DW 6] and the curriculum aims to provide clear explanations. Then the curriculum goes on to show users on how to perform analysis on the paradata recorded to the log. These paradata tutorials are meant not only to help orient CNP data workers to these advanced features, but to help them position themselves as dataset auteurs. For example, one of the videos covers how to visualize user contributions over time, while another shows how to plot the range of edits over time, which, in many cases, can demonstrate the depth of experience of a dataset contributor.

Focus Group: D4D Dataset and Data Work Terminology

Motivation. 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 workers who perform the labor that makes AI systems possible. 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. In our user testing, we noticed something: even though several participants were familiar with the concepts that were covered in the Datasheets 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 Datasheets project–rather, it follows a budding genre of work that extends the Datasheets project for specific domains and practices. D4D has proven useful as a tool for critical thinking about the data used to train AI systems [385]. For example, datasheets have been adapted for Speech Language Technologies [390] and for NLP [391]. Our investigation is centered on the vernacular language that data workers use to describe the datasets they work on and data techniques they perform. 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).

Methods. 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 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, asked them how they felt about the equivalent D4D term (here, "instance"). The third category of question was related to higher level concerns about data, for example, about terminology data that could provoke emotional responses, and whether D4D's use of "anxiety" felt sufficient.

Contextualized D4D Discussion Guide. These questions are split into two portions: the first relates to applied work, or the experience of performing the data labor, while the second pertains to the data worker(s)'s outlook and perspective on performing data work. While this guide uses specific language relating to DataWorks and prior projects, these questions can be easily adapted to other organizations with appropriate swaps for organization-specific terms.

Applied Questions

- How do you describe datasets when they first come into DataWorks, based on what kind of work you will be doing to that dataset? What kinds of words do you use – dirty, disorganized, un-standardized?
- What do you think of when you hear "raw data"?
- How do you describe the work you do at DataWorks, in terms of what happens to the data? Are you cleaning it? Organizing it? Standardizing it? Refining it?
- What do you think of terms like "preprocessing", "cleaning", "labeling"? Do they fully encompass what you do? Or do you feel like other terms would be more appropriate?
- Do you feel like your work is often data collection?
- How do you refer to a single unit of data? A single unit of the dataset that you're working on, e.g. ...? Do you use a word other than "instance"?
- When you're working with data, how do you describe a unit (instance) of data that you might not expect to see or that seems wrong or faulty to you? What about when there are two or more entries in a dataset with the same information? Are "errors", "sources of noise", "redundancies" meaningful?

Theoretical Questions

- Can you imagine ways in which you work with data (cleaning it, for example), that might impact future users? How do you anticipate and address this? Documentation practices?
- When you provide information to a client on how to work with / use a dataset that you've finished working on, what do you call it? The directions–are they a "read me"? Or does it take another form?

- Do you usually know who "created" a dataset (except when you are collecting data directly, for something like GreenLight?) How do you define your own role in terms of data creation and use process? Do you feel like a curator?
- How do you decide if you feel comfortable working on a project? (both at organizational level and then on a personal level)
- Thinking about a time that you were working with data that was difficult on an emotional level, or made you uncomfortable, or anything negatively-out-of-the-ordinary, what emotions were you feeling? How would you want this data to be described, or these entries that provoked this emotion to be described?
- What is private data? What about "sensitive" data? What kind of data do you think could be used to discriminate or unfairly represent a given human population?
- What are some examples of "confidential" data that you've worked with at DataWorks?

Findings. This analysis is split into participants' insights regarding first the applied questions, and, second, the most holistic, or theoretical, reflections on completing data labor.

Applied. The language that the Data Fellows used to describe the data they received from clients generally revolved around the term "raw," though with at least one Fellow pointing out that there is no such theoretical concept as "raw" data, but rather that the data format was straight-from-the-source, or had yet to be cleaned. Here, the use of "raw" as compared to its use in academic discussion, e.g., by Geoff Bowker [273, 274], conflicts with those of practitioners: while the Fellows are aware of the theoretical concept of "raw" data, it is still an appropriate term to describe an intermediate data process, where the client is the point of dataset genesis. This use of "raw" applies similarly, in the Fellows' perspective, to data scraped from a webpage that remains "straight from the source," "organic,"

or "untouched and unchanged." The work that then takes places to transform this dataset from being "raw" to being actionable by the client may include things like correcting misspellings, lack of categorization, improper capitalization, and unifying ways of spelling–all things that fundamentally change the shape of the dataset but are unanimously lumped into the "preprocessing" stage in D4D terminology. The term "preprocessing" itself was actually used by one Fellow to describe working with data shared by a client that was not the most up-to-date version version. On that particular project, when the Fellows' shared their work with the client, the client was frustrated to find it was not the final version of the dataset they wanted to have cleaned. Here, "preprocessing" was the term used to describe the client's work on the dataset, specifically, verifying that the dataset version was correct.

Similarly, language to describe surprising or questionable data transformations was based on either colloquial terms or the actions the Fellows would need to take to remediate it. Such a data point was described as an "outlier" or simply "doing too much". Terms for problematic data points (such as "errors", "sources of noise," or "redundancies") were swapped for the action that a Fellow would take–e.g., "flagging" ("I would flag that" or "We flag it") or "hopping on a [Microsoft] Team's call". The unit of data in question might be described as a "datum" but could also be a "cell" or a "line," corresponding to the type of concern raised.

Theoretical. When considering the data labor that the team performs as a whole, the Fellows typically know who their downstream users were (their clients) and are able to communicate with them directly, through presentations and data manuals that are tailored to their particular needs (for example, in response to specific client concerns or questions). This makes imagining the range of downstream users, beyond the client, more difficult, because the focus is on the client. This raises an interesting question in conjunction with workplace autonomy and consciousness; if AI practitioners cannot anticipate every potential use of a dataset, in part due to the constantly evolving nature of the field, how can data workers be assured of the limits of their participation, beyond immediate downstream

users? This tension–immediate vs eventual downstream users–also suggests a disconnect in the way D4D currently handles anticipated use, vs the way that data workers actually encounter it.

The idea of dataset domains or topics that could produce "anxiety," as described in D4D, also challenged participants' notions of how decision making in data labor takes place, specifically at DataWorks. Given that DataWorks does have a unique structure, conversations about consenting to labor do look different than at other data service providers. However, what is likely more generalizable, is how the Fellows' challenges the idea of any one source of anxiety as being universal. The Fellows recounted various projects that touched on topics that may have been sensitive for one or more team-members, but were not universally held as such. The mechanisms used to navigate such tensions tended to be based on open communication within the team, and had less to do with any one label or warning prescribed by the client on the dataset, based on domain subjects pertaining to that dataset.

The Fellows were unsure if the term "anxiety" was sufficient to describe the range of emotions that they felt working on sensitive subjects; while they saw connections between things they had experienced and the term itself, they pointed out other phrases, including datasets that might provoke "worry" or "a reaction". Further, the team shared mechanisms that supported them in working on sensitive topics, including taking breaks and other physical movement practices, such as to release muscle tension.

The final term in question was the concept of "sensitive" data–in the D4D term, meant to describe data that could be used to discriminate or unfairly represent a given human population–was familiar, though not by name. The Fellows routinely work with civic and non-profit organizations, whose data often quickly reflects individuals or groups of individuals who are being helped or supported by the client organization. Unsurprisingly, the idea of "confidential" data thus revolved around the idea of the individual whom the data depicted. This is in large part a reflection of the general orientation of DataWorks clients

(i.e., civic and non-profit) as opposed to commercial practice, which might see things like locations or events as carrying the same designation.

The contextualized definition of "sensitive" raises another domain area question: what kinds of data should be considered universally sensitive, if any? How should individual organizations, or data workers, without vast experience in downstream AI work predict or ascertain whether they are working with potentially sensitive material? There are perhaps obvious examples of sensitive data–the Fellows identified children, the elderly, or other vulnerable human populations, as well as any individual's face or name as such–but what happens when downstream AI model can (if imperfectly) infer characteristics like race or ethnicity from zip code or surname [392, 393]? This is a key consideration for dataset contextualization, considering that data annotators are rarely informed of the final projects or systems that are based on their labor [149, 22, 91], with the Fellows concurring that this was also their experience with only a few exceptions.

In short, this focus group highlights 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 AI parlance, it functions as a kind of *Newspeak* through which verbosity obscures facts. As Slava Gerovich has shown [219], 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 labors. Specifically, the distance that developers may be putting between themselves and the vast array of professionals who make their labor possible.

4.3.5 Next steps

The Datum Fieldnotes project has two avenues of continued work. First, there are technical changes still to be made, resulting in a Version 3.0 of the tool. Second, the results of the focus group conversation with the DataWorks Data Fellows about their understandings of D4D terminology highlighted a number of challenges to the adoption of D4D by data workers directly. These topics are discussed in turn in this section.

Design Insights for Version 3.0

First, the variance data in work backgrounds of our participants suggests the need to develop tools for CNPs that have both graphical interfaces (such as the Datum Fieldnotes' sidebar interface) for new to data work, to minimize the burden of adoption, particularly in the case of domain experts who are secondarily performing data work [374, 267].

Second, based on the range of data types our participants were working with—ranging from textual, to numerical, and manual calculations vs imports from tools and sensors—we should consider ways that spreadsheets and their add-ons, as a sustainable software for CNPs, can be adapted to support annotation of additional kinds of data, for example, images.

Third, the security and privacy of CNP data remains a space rife with tensions between different kinds of data and data workers. For some of the participants we talked to, being able to anonymize or lock down their datasets (while still making them usable and annotateable within the organization) was a primary concern, particularly for organizations with clients experiencing conditions with social stigma (e.g., homelessness or health conditions). Other participants wanted to be able to share out their annotations (as public paradata)

and wanted to know how Datum Fieldnotes would or would not support that functionality. Additional work is needed on our part to develop a version of the tool that will satisfy the needs across the security and privacy spectrum presented by our participants, with our tool version two presenting some initial improvements. While Datum Fieldnotes is one example of the privacy implications of CNP data work, our observations highlight a key structural problem: CNP data workers need custom, value-sensitive software and tools, but do not have the budgets to support it.

Adapting D4D to Data Works' Context and Vernacular

There are two main contributors to the discordance between data workers' vernacular terminology and that used by D4D, which represents the language of AI practitioners more broadly. First, the language used by AI practitioners abstracts over the nuance that data workers see. In particular, there is a lack of understanding about not only the depth of labor that data workers perform, but also about the discretionary choice involved in the data transformation(s) that take place. Given this discrepancies, it is not only the case that a D4D that is accessible to data workers should include more room for these discretionary choices, but also that AI practitioners likely need information to digest this new information.

Second, the two groups' respective notion of future use and downstream users appears potentially distinct. In many cases, data workers (both at DataWorks and other organizations) may be aware, at best, of the use of their labor by an immediate or a direct client. However, farther downstream users, e.g., other AI system developers, or even non-AI dataset re-use. This topic has come up previously in literature, mostly in the form of trusts or data banks for data workers, such that when their existing data labor are re-purposed (or re-sold) for a new project or application that are compensated [95, 149]. This discrepancy is exemplified by the way the Fellows described putting together contextualization materials for clients, which took the form of presentations or write-ups that were tailored to the needs of the requesting client.

Subsequently, there are two next stages of work that are necessary to making D4D (or similar data transparency artifacts) an active site of collaboration between data workers and AI system designers and developers. First, the terminology used should be expanded to reflect the vernacular(s) used by data workers, and namely, the emphasis that data workers place on making sense of things like "sensitive" content in context of a particular dataset. Second, the lingering downstream user should be the source of discussion between the two parties (data workers and AI system designers & developers). If data workers are to perform the dataset auditing that they are best-positioned for, then that auditing must be done with a sense of who might use a given dataset and how they might use it.

Further, the question of workplace autonomy and consent to labor remains an indivisible element of understanding downstream use and users. In the development of technologies that might be used, for example, in surveillance or violence, the link between the data work needed and the end product may not always be immediately clear to individual data workers [394]. Of course, in the case of data, as system development evolves–including the technical paradigms for what systems can even be constructed–there will always be an element of ambiguity about what happens downstream [395]. However, what is more clear, is that AI system designers and developers should undertake (more strenuous) effort to onboard data workers to the project's goals and motivations, as well as contexts, so that data workers can complete their audits with an eye towards what is anticipated, as well as what might be yet undetermined future use.

4.3.6 Limitations and conclusion

This work has two limitations. First, as we have existing relationships with our participants, we are aware that they might have felt compelled to provide us with positive feedback. In order to preempt this, we structured all recruiting materials and interview questions to make it clear we wanted to elicit critical feedback and we wanted to make the tool usable for our community partners and made a point of asking for features that were uncomfortable or

difficult to use, as well as what additional features would be needed to make the tool useful and practical. Second, as with any qualitative work and user testing, our sample size is not exhaustive and does not reflect experiences; however, given the range of data practitioners (organization type) and format of data work, we spoke with, our insights point towards multiple avenues of future work, particularly related to data privacy and sharing tools for CNP data work.

Our contribution is the design and deployment of a Google Sheets add-on to offload the documentation of individual data transactions and re-position data workers as dataset auditors. Complemented by a suite of user testing, we demonstrate the difficulty of designing with and for CNP data workers and identify underappreciated tensions in the design of new data tools for an often-overlooked, but critical, population of data workers. There are two major contributions of Datum Fieldnotes to the way data workers create and process spreadsheet-based datasets. First, the tool allows CNP data workers to easily append their observations, annotations, and reflections in a way that is both easy and records the material in a CSV so that it can travel with the dataset itself. Second, it supports CNP data workers in learning new spreadsheet skills and enhances dataset documentation, by automatically recording per-datum changes to the spreadsheet, so changes can be easily traced. This late-breaking-work focuses on the design and development of the tool, from the first version to the second version that was informed by a set of user testing session we ran with twelve CNP data work professionals.

4.4 Chapter Conclusion

This chapter is study of what perspective brings to data work, or how datasets changes when we recognize them as products of a sociotechnical development process, and given data workers both the tools and the space to interrogate them as such. What we get are datasets that are actually grounded in their origins and that recognize and respect the phenomena that they present. This means the analysis and systems based on these safe (as

contextualized) datasets have a much better shot at behaving as we'd like, or as we anticipate.

This chapter is grounded in the case study of DataWorks. First, I created and taught a critical data literacy workshop at DataWorks to help the Fellows translate their existing critical perspective into their data work. The Fellows pulled on that knowledge when they, as described in the subsequent case study, objected to a project brought to DataWorks, in which they were asked to annotate a dataset relating to AAVE. Their perspective and critical take on the project, partnered with the unique structure of DataWorks, meant that they were both treated as collaborators and paid while acting as essentially consultants on the project. Their critically was not just a means for contesting the work as it existed; instead, they pulled on their knowledge of ML and AI systems to design a new interaction of the task and reiterate their positionalities in ways that definitely enhanced the project.

To formalize this practice, I share the design and development of a data tool, *Datum Fieldnotes*, a Google Sheets add-on. Datum Fieldnotes has two purposes. First, it records the annotations that data workers make on the Sheets project (specifically related to datum concerns, questions, corrections, or observations) and makes it easy to share them out. Second, it automates per-datum documentation via the Log functionality, creating a workable body of paradata that can be used to better contextualize and understand the dataset at large. Between these two affordances, data workers using Datum Fieldnotes can additionally use them as a way to demonstrate their expertise in, and contribution to, a given dataset. This can be used to demonstrate both personal expertise and skills development (for the purposes of professional promotion) as well as authority when crafting and sharing the data narrative(s) that emerge from the dataset.

CHAPTER 5
DIVERSIFYING PERSPECTIVES AT SCALE

The prior chapter is devoted to demonstrating what engaging with, and embracing, individual perspectives can do for datasets. In this chapter, I explore methods to take these findings beyond DataWorks. However, given the inherent differences in different data work sites, I adhere closely to the idea of the *cosmolocal*, to an approach to scaling-up that acknowledges and respects the difference between different sites of production, while still contributing a body of global knowledge to be shared between all sites.

The goal of this chapter is to make sense of how the conditions at DataWorks, both in interface between workers and clients, as well as training courses and workplace setting—which result in data workers taking on the role of dataset auditors—may or may not scale up to platform-mediated data work and data workers on those platforms. This chapter comprises two projects which are focused on re-balancing the power dynamics between data workers and requesters of their labors, without expecting significant help (in either direction) from the platforms mediating that labor exchange.

First, I offer help to academic requesters of platform-mediated data work create more pro-social workplace conditions, with a guide to what is actually needed to create the kind of labor setting as experienced at DataWorks. The goal here is not to make sites like Amazon MTurk a DataWorks clone, but rather to work within the existing infrastructure of the platform in order to reposition the data workers (Turkers, in this example) as requesters' collaborators.

Second, I share the results of broad survey of requesters specifically employing platform labor for data tasks used to train or refine AI systems. Targeting the way that requesters understand—both interpersonally and functionally—the data work professionals who fulfill their task requests, this work is an examination of what is needed to make collaboration

possible. Critically we find that requesters struggle with the idea that platform data workers both represent the general public (and subsequently, are employed, as a way to get a public auditing of a system, though it was never described as such) but do not offer the labor conditions needed for data workers provide such labor.

Together, the two pieces of work included in this chapter point out current incompatibility with the data worker as auditor model, despite the fact that, based on the second work introduced in this chapter, many requesters are looking for such a thing. In the final chapter of this dissertation, I explore avenues for building on this work in the future, to show how some of the conditions needed to develop the style of collaborative relationship necessary for the data worker-as-auditor model can be developed in-place, within the infrastructure of these platforms.

5.1 Designing Pro-social Data Work Experiences

The use of crowd-working platforms (e.g., Amazon Mechanical Turk [AMT], Clickworker [CW], and Microworker [MW]) to build large-scale machine learning training datasets is both widely accepted and prevalent in academic research practice [139]. What is also prevalent is academic researchers abusing the power imbalance of these platforms, leaving digital pieceworkers [23] at the mercy of their academic requesters [275]. Given the dehumanizing title workers are given—they are only a subunit of the larger "crowd" [396] – these practices are not surprising; indeed, while there are bad actors, the distributed nature of the platforms may cause requesters to appear to enact these malpractices out of misunderstanding more than malintent. For this purpose, we propose a cover sheet describing precise hiring and employment practices of academic collaborators (or "crowd collaborators"[1]) engaged through crowd-working platforms. The cover sheet is to be submitted with the publication of the resulting work by researchers in academic venues. The design of the

[1]While we cannot find the original use of this term, we are sure it has been used in prior work and the original author(s) should be credited. Our use of the term pulls from the collaborative nature of digital piecework workers and academic requesters who use their services (described in [397, 398, 399]; our intention is to highlight the value of the work that digital pieceworkers perform and highlight their contributions.

cover sheet is inspired by the Datasets for Datasheets project [288], to be required by academic venues accepting the results of crowd collaborator labor, namely machine learning training datasets. As Hießl argues, crowd collaborators do not have traditional employment contracts to rely on and that a new form of contract must be developed to address the complexity of digital piecework [400]; we present this cover sheet as a first step in that direction. By surfacing this information at the institutional level we hope to 1) inform requesters of the best practices if they are unaware, and 2) certify respectful treatment of crowd collaborators, especially given the calls to substitute digital piecework for jobs lost in the face automation [88, 401].

Our intervention[2] centers academic requesters for two reasons. First, we choose to highlight the role of requesters in these platforms as the power balance is inevitably shifted away from those performing the labor to those providing it, due to the oversupply of labor and undersupply of available tasks [400, 201]. Previous interventions have sought to reform the digital piecework labor system from other angles. One of the most well-known, the TurkOpticon project [79], helps crowd collaborators source much-needed information about the requesters and tasks they encounter. However, TurkOpticon's founders realized it became a permanent, relied upon feature of the ecosystem rather than forcing Amazon's hand to create permanent, platform-implemented safeguards for workers [80]. Similar efforts to improve the platform from the worker side include the Crowd-Worker plugin [402], Crowd Guilds to unite workers [82], and a worker-owned cooperative model alternative [403]. These innovations either beneficially augmented the experience of workers or proposed alternatives; however, they require external funding and, in some cases, forgoing immediate profits for long-term vested interests, which is not an option for workers who need immediate payout [139]. Further, there may exist inequalities in the way different crowd collaborators are rated, where applicable [120]. Similarly, while unions promote higher income and feeling of community between workers [123], these digital work plat-

[2]Work done in collaboration with Justin Booker, Christa Davoll, Jessica Hill, Venise Ivey, Carl DiSalvo, Benjamin Rydal Shapiro, and Betsy DiSalvo, which appeared at alt.chi in 2022.

forms sometimes act against them, as in the case of Amazon Mechanical Turk banning the account(s) of small, collaborative groups [22].

The *We Are Dynamo* project [404], which unites workers to achieve collective action, has provided among numerous outcomes a best practices guide for academic requesters, however there's no system to enforce academic requesters' adherence to them. Whiting et al. [94] inspired our decision to shift the focus to requesters as they trust workers to report the time they took to complete the task (or a reasonable approximation) in order to guarantee proper payment for work performed. Requesters hold a great deal of power over crowd collaborators, as does the platform. However, the latter seems impervious to improvement in the short term, as described by the operators of TurkOpticon, whose improvements to the AMT platform for workers were not adopted by AMT itself [80]. We hope to combine the common issues surfaced by these efforts and provide a way to operationalize their findings and concerns by mandating compliance at the institutional level, similar to the IRB process for human subjects. Our goal is to extend existing knowledge about what a fair requester-worker dynamic looks like into a formal reporting system to create a more just and respectful workplace for crowd collaborators.

Second, we highlight the role of requesters from inside the Academy. As requesters, academics and our industry collaborators—as highlighted by Scheuerman et al.'s study of computer vision researchers—are failing to meet basic standards (e.g., clear standards for terms of employment) for fair digital piecework practices [275] despite the popularity of such platforms [405]. Further, as we continue to confront the biases embedded in our research designs and products with regards to data, we must acknowledge that—in many cases—they are the result of our own oversight and overly-generalized practices rather than the fault of our crowd collaborators [406, 407, 51]. Along with prescient data about the terms of employment, we ask that requesters engage in a reflection of what values or experiences are reflected in the data work they request.

This work presents the cover sheet as a specific contribution, but also seeks to engage

in dialogue with the larger ACM academic community to evolve the notion of cover sheets and other related ideas. We do not seek to end the practice of sourcing digital piecework through crowd-work platforms. We recognize both the research and employment opportunities that these platforms provide, especially with respect to workers who may have preoccupying care-giving tasks, difficulty travelling to a workplace, or face discrimination in the workplace. Rather, we hope to institute a more sustainable practice that engages crowd workers as collaborators, acknowledging both the injustices that academic requesters have perpetrated on crowd workers and the changing nature of labor in the face of automation.

5.1.1 Methodology

Our methodology for constructing the cover sheet is based on a mix of first-hand experiences in the reflective style of [408] as well as direct observation by professional data workers, and finally previous findings by academic and crowd worker community bodies (e.g., We are Dynamo [404]). The first version of the cover sheet was designed based on the first author's observations from the experiences above and existing literature, structured along the comparison axes the data workers highlighted in 5.1. The data workers then provided feedback during a 1-hour session which resulted in the second version of the cover sheet shared in this chapter. In the explanation for the different pieces of information, the data workers are quoted directly or summarized in brief from the research records collected during the engagements listed below.

5.1.2 Setting

The second through fifth authors are employees of DataWorks, a work training program for developing the skills of a mid-skill data worker incubated in the Georgia Tech College of Computing. The program aims to broaden participation in the everyday work of data collection, cleaning, and basic analysis. DataWorks' employees (the "data workers") are people from economically disadvantaged neighborhoods and underrepresented groups in comput-

ing and DataWorks aims to assist them in finding solid, middle-class jobs in data work. In many ways, DataWorks functions as an alternative to more classical digital piecework platforms like AMT, CW, and MW. The data workers engage with a variety of client projects, a portion of which resembles classic image recognition or natural language processing sentiment analysis tasks that would be found on the aforementioned platforms, along with more typical spreadsheet-based projects. Unlike those platforms, however, employees of DataWorks are full employees of Georgia Tech and therefore receive a competitive hourly wage ($17.35), health care (USA-specific), other fringe benefits, organization-provided computers and workspace and work a 40-hour week with regular hours.

Further, the data workers have extensive input on client projects and engage in dialogue directly with the client, including—depending on the project—initial training sessions, clarification questions, and project presentation at the conclusion. DataWorks' client projects are longer term than discrete digital piecework tasks; for example, the data workers identified and summarized the events of close to 900 cartoons for a single requester. The data workers have a skillset that is—with regards to this kind of digital piecework—therefore comparable to an experienced, professional worker on the more classical platforms (e.g., AMT, CW, MC). While the data workers are not full-time crowd collaborators, their expertise and experience play an important role: they are aware of alternative structures to classical platform work – via DataWorks—and, given the research setting of the workshop, they can take the time to ideate and critique. While investigations that center experienced, professional crowd workers are of immense importance, we believe that adding the accounts of the data workers is an important contribution.

5.1.3 Construction

Over the course of seven days in the summer of 2021, the data workers, along with the first author, engaged in a reflective workshop to compare the experience of working for DataWorks with that of a crowd collaborator on three digital piecework platforms: AMT,

CW, and MW. Other platforms were initially investigated but the workers were not able to complete work on the platforms due to the location requirements for Sama (formerly Samasource), full-time requirements for LeadGenius (formerly MobileWorks [111]), and lack of available tasks for Appen (acquired the former Figure Eight platform). The workshop took place under the auspices of university IRB approval and the data workers were paid their normal hourly wage while engaging in the workshop. The workshops enable us as researchers to better understand work practices and provide the workers with domain-specific skills and business practice.

The workshop was intended to identify what aspects of employment for digital piecework DataWorks was getting right and which aspects the organization could improve. The workshop was open-ended and began during the fourth week of a 10-week summer tutorial course designed and facilitated by the first author on the politics of data and key data cleaning and standardization skills. The point of the workshop was to directly engage with and observe alternative employment systems for digital pieceworkers and compare and contrast experiences on those platforms through discussion. Unlike other experiential work on digital piecework platforms (direct observation by researchers or interviews with, or observations of, crowd collaborators on those platforms), the data workers have the professional experience of being digital pieceworkers and given a lack of time pressure, were able to reflect on their experience and brainstorm alternatives. The data workers' impressions were collected through four kinds of engagements:

1. The data workers engaged individually with the platform (to mimic the isolated nature of digital piecework), including signing up and working on the three platforms. The data workers recorded their impressions on shared and individual note-taking documents. Duration = 7 hours, broken into multiple sessions. The data workers kept shared and individual documents of running notes and discussed their experiences with the first author in their regular interviews (see item four of this list).

2. The data workers worked communally on a given task, with one operating a computer from which the task was projected on a large screen, and all team members engaged together on the task, more akin to the collaborative setting in which the data workers usually operate. During this session, the data workers trialed the TurkOpticon browser add-on and were introduced to other AMT community resources, such as TurkerView forum. At the end of the session, the data workers compared their experiences on the three platforms (see Fig. 1). Duration = 90 minutes. The session was recorded and the first author took notes.

3. The data workers described their experiences to the sixth author, who was unfamiliar with the platforms, in a 90-minute session, intermittently working on tasks together during that time to demonstrate their reflections. Duration = 90 minutes. The sixth author took notes and the data workers expressed their recollection in their regular interviews (discussed below).

4. Finally, throughout the workshop, the data workers met individually with the first author to reflect on their experiences in semi-structured interviews. Duration = roughly 15 minutes per worker per week, for three weeks. The first author audio recorded interviews and took notes during the sessions.

In all, the data workers accrued more than 10 hours each of experience on the three platforms, with the majority on AMT, through two 90-minute tutorial sessions and 7 hours of independent work spread over multiple days. Some workers did not attend all workshop sessions due to other conflicts, but all workers completed at least 10 hours of experience on the platforms combined. Where possible, the data workers never "cashed out" payment. Because they were forced to "cash out" to register for some platforms they earned $8.98, which went towards snacks for the DataWorks office. The data workers were not required to provide their personal information in order to use the digital pieceworker platforms.

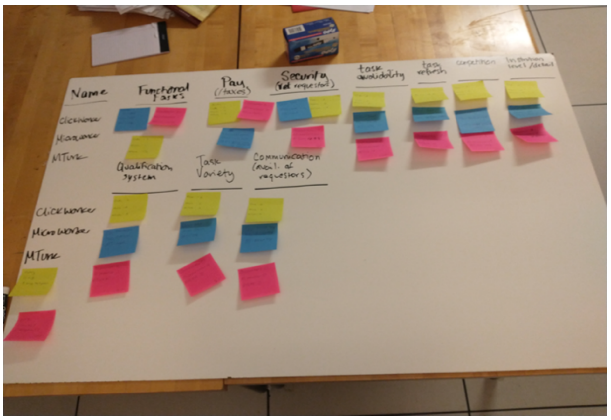


Figure 5.1: A comparison chart of the three platforms by three of the data workers. As the sticky notes are handwritten, a typed version is available for download at `https://annabelrothschild.com/documents/alt-chi-22/Fig1-text-translation.docx`

5.1.4 Researcher positionality

This work was done as part of collaboration with the data workers[3] whose insights and backgrounds significantly contributed to the design and development of the coversheet described in this chapter. Additionally, the first author, who interacted most with the data workers to design and development the coversheet, has extensive experience (spanning four years) as both a worker and requester on AMT. These experiences include early experiences working at Wellesley College, first with Dr. Eni Mustafaraj then with Dr. Ada Lerner, working on AMT in order to understand how human intelligence tasks (HITs) should be developed and later as requester for both quantitative and qualitative tasks. Practices developed in the Cred – motivated by [409]—and Security & Privacy labs (respectively) with both PIs and other student researchers inform the provided practical examples of how requesters can, among other things, calibrate payment and employ respectful demographic questionnaires. These experiences likewise significantly influenced the design and development of the coversheet described in this chapter, for example, to inform portions of the guidelines that refer specifically to the ways in which requesters can structure and reflect

[3]The full author list of the version of this chapter published at alt.chi 2022: Annabel Rothschild, Justin Booker, Christa Davoll, Jessica Hill, Venise Ivey, Carl DiSalvo, Benjamin Rydal Shapiro, and Betsy DiSalvo. Authors 2-5 contributed in their capacity as DataWorks Data Wranglers.

on their digital piecework tasks.

5.1.5 Cover sheet items

The cover sheet[4] is meant to surface information about the conditions of collaborator employment and build an ecosystem that is more respectful of crowd collaborators. Features of the AMT platform are often used as examples because the platform is one of the more robust and popular with the computing community; however, there are identical features and mechanisms on other platforms. Some of this information (e.g., payment) has often been reported but not uniformly [275]. We do not think filling out the cover sheet should take more than an extra hour for the research team; however, the tasks required (e.g., follow up with accepted or rejected crowd collaborators) may require extra time during the experimental phase of the research. We argue below that each of these additional tasks has a meaningful effect for crowd collaborators and should be required as such. As we describe in 4 (Future directions), we will be confirming this (and iterating on, as necessary) in a field experiment. An additional practice that academic requesters should consider is monitoring their reputation (via their profile) on forums like TurkerView[5] to proactively catch problems crowd collaborators encounter, a practice employed by the Cred Lab.

We build off the recommendations of the Dynamo guidelines [404] and use the ± symbol to indicate reiteration and expansion of recommendations developed by the We Are Dynamo movement.

1. **Basic information**. This information should be described to help the academic community receiving the contributions of the crowd collaborators assess the context of the task.

    1. *The platform used (e.g., AMT).* For reproducibility – described further

[4] An example of the cover sheet as a fillable PDF, along with a completed example cover sheet, are available at: `https://annabelrothschild.com/projects/alt.chi-22/pro-social_crowd_collaborator_recruitment_guidelines`

[5] `https://turkerview.com/`

in [410]—and to assess per platform specific features, some of which are discussed in [411].

2. *The requester name used to post the task* ±. Providing clear, factual information in the requester name (e.g., Prof. X, University Y Lab Z) can help crowd collaborators understand who they are working for and track the progress of individual tasks in the post-submission phase. The former allows crowd collaborators to help describe the nature of their work and recognizes them as collaborators, while the latter helps collaborators track their salary and address concerns. As data worker stated with regards to platform work, "*Can you even use this...can you put it on your resume, is it respected work*?" Acknowledging the skilled labor required to complete HITs for ML training data, requesters should allow their collaborators to signal their expertise. In addition, if an academic requester has a faulty task and fails to state (or misstates) follow up information, the researcher can be found via their public institutional profile online.

3. *The full HIT name and short description with task category* ±. More information readily available to crowd collaborators allows them to cut down on the significant labor of sifting through available work[402, 139, 94]. The data workers also highlighted the importance of knowing the task category (e.g., image "tagging" for recognition) in conventional crowd collaborator language (e.g., "chat with a bot" for NLP conversational work).

4. *Contact information given to the crowd collaborators and designated team member(s) who monitored inbox* ±. The contact information (e.g., email) provided to crowd collaborators should be made available to ensure that it is accessible. Designated research team mem-

ber(s) should be "on call" to monitor the contact inbox to ensure that crowd collaborators can receive follow up within a reasonable time frame; the data workers' consensus was 24 hours was appropriate and this number should be confirmed in future work. For example, in Drs. Mustafaraj and Lerner's labs, HITs were posted with a contact email address that would automatically forward to the PI and research assistants on the project, or a lab address that research assistants running studies could access and the PI could review; the individual student(s) running the experiment are then responsible for monitoring that email address. Particularly when apprenticing researchers are involved (i.e., students) who may be new to running experiments on crowd labor platforms, a more experienced member of the team can ensure that the apprentices are engaging properly with crowd collaborator inquiries.

5. *IRB consent form, if applicable* ±. For archival purposes; can be attached to the cover sheet as supplemental material to ensure coherence with.

6. *Warnings provided about potentially sensitive activities or topics*. Crowd collaborators should be given enough information about sensitive topics in a task so they can make an informed decision about accepting the HIT without having to scroll through multiple warning screens—or worse—be forced to abandon the HIT partway. This respects their time and does not affect their return rate, which can be used as a collaborator qualification on AMT, for example.

7. *Time(s) of day and day(s) of week HIT posted, including the number of HITs posted in (each/the) batch*. This information should be provided to gauge potential population bias or impacts on crowd collaborator lives. For example, requesters should respect local time zones—and,

if hoping to achieve a global reach, should post their tasks at times that are conducive to collaborators in those locations [22, 88].

2. **Crowd collaborator treatment**. This section addresses concerns that directly relate to the treatment of employed crowd collaborators, including—but not limited to—fair compensation, ensuring a right to privacy and security, and structuring a HIT as accessible as possible.

    1. *Terms of employment*.

        1. *Number of collaborators desired and proposed payment per crowd collaborators, along with any later bonuses paid out.* The number of collaborators refers to the number of distinct individuals who complete the tasks to determine the diversity of the collaborator population who performed a certain HIT, which may have implications for the use of a given dataset, given the subjectivity of the work. Further, requesters should state how much they intended to pay collaborators per HIT and how they arrived at that number as described in [412]. Suggested methods include interacting with the locality-respecting calculator built by Sinders [279] and the one-line of code used to guarantee a $15/hr wage[6] built by Whiting et al. [94]. Additional bonuses should be described; e.g. via Whiting et al.'s mechanism or as part of total compensation used to follow up with individual crowd collaborators (AMT, for example, requires a minimum 1 cent USD for the "bonus" payment mechanism which can be used to communicate with collaborators after they have finished a HIT). The Dynamo guidelines for aca-

[6] As of 11/22/2021, $15 USD in the following highly populous countries = Chinese Yuan: 95.79; Indian Rupee: 1116.05; Indonesian Rupiah: 213745.50; Pakistan Rupee: 2628.63; Brazilian real: 83.78; Nigerian Naira: 6172.72

demic requesters state that $0.10 USD per minute is considered an effective pay floor and that "tasks paying less than $0.10 a minute are likely to tap into a highly vulnerable work pool and constitutes coercion." [413] While many requesters operating out of the United States may consider applying their state or district minimum wage, consider there are a wide range of minimum wages in the US (from $7.25 to $15 at the time of writing). Without asking the collaborator about their location (with IP address being a poor proxy as Whiting et al. describe [94]) it is difficult to ascertain proper minimum wage; for this reason, Whiting et al 2019 default to $15 per hour. Further, d'Eon et al. describe the mutual benefice of fair wages and how they might be calibrated [414].

2. *Number of crowd collaborators accepted and percent accepted rate; number of crowd collaborators rejected and percent rejection rate.* A high rejection rate (context specific, but generally more than 10%) can indicate a faculty HIT; for example, a mechanical issue or a lack of clear instructions. Further, rejection without clear rationale can indicate that collaborators were unfairly rejected so the requester could get more labor for less compensation. A high rejection rate should be explained and follow up action should be described, such as a soft-reject (compensate collaborators who did the task correctly to their understanding but incorrectly for the purposes of the researcher; in this case, the researcher "accepts" the HIT and pays for the work, given that it was their mistake). For clear-cut tasks, rejection rates are expected to be low. The data workers described engaging in a task that required copying and pasting the results of a Google search query ranking about which

there was no ambiguity; they were rejected either without rationale or a confusing "*nice work!*" message which indicated either malintent by the requester or accidental action (the requester never responded to follow up messages from the data workers).

3. *Criteria for rejection (list)* ±. Pursuant to immediately preceding item, requesters should summarize criteria for rejection after reviewing multiple rejection-worthy entries and follow up with individual crowd collaborators to communicate cause for rejection. This assures the collaborator that the rejection was legitimate (e.g., a spam entry or failing reasonable "attention checks") if there was a mistake on the part of the requester (seeming accidental rejection) provides the collaborator with a clear way to request clarification. Where possible, rejection rationale should be communicated as "fruitful feedback" [415].

4. *Follow-up method to communicate rejection/acceptance for each crowd collaborator.* Pursuant to the preceding two items, crowd collaborators often site lack of reasonable follow-up and communication from requesters are a major problem [97, 123]. All follow up should include the HIT name and requester name in communication. Before posting the task, researchers should assess the follow-up mechanisms of the platform and if they must collect additional information to engage in follow-up add that to their task with clear rationale for doing so and allow given collaborators the opportunity to opt-out (in case they do not want to provide a mechanism for follow-up out of privacy concerns). Any future contact information collected must be allowable by the platform Terms of Service.

5. *If disallowing multiple submissions by a given crowd collaborators, state mechanism used to do so* ±. A platform's "blocking" feature should not be used as it limits the future work available to a crowd collaborator by disbarring them from future, unrelated tasks from the same requester. Instead, on AMT for example, requesters should make use of the "qualification" mechanism to disbar multiple entries for a single HIT. If a qualification mechanism is used, make the purpose of qualification the qualification name (e.g., "July2021StudyNoMultipleSubmission") to help crowd collaborators track HITs completed and reduce ambiguity around random qualifications [416].

6. *List any required collaborator qualifications and rationale for them*. Extensive use of qualifications limits both the pool of available crowd collaborators and the work available to crowd collaborators. Previous research has shown that not all distinctions are necessarily reliable metrics [417]. Given the over-subscription of workers compared to the number of HITs available [400], requesters should be conscientious to extend work to all legitimately qualified collaborators, pursuant to task type characteristics. Given the prevalence of unfair rejections, requesters should be sensitive to using approved HITs as a qualification metric; collaborators who are unfairly rejected must then work a high number of HITs correctly to fix their acceptance ratio, which can force them into low-paying and exploitative work.

7. *State any pre-test tasks.* Sometimes HITs require particular qualifications that must be described by the individual collaborator. If requesters are seeking specific demographic characteristics, for ex-

ample, they should consider using a platform that caters more directly to that need—for example, Prolific appears to be one such, but individual requesters should confirm this. If collaborators will potentially be disbarred from a task given their pre-test results, they should still be compensated for their time as they produced labor and information (if incorrect) for the requester. Malicious requesters may require extensive pre-test information that allows them to get the majority of their HIT done despite rejecting most (or all) crowd collaborators. Rejected crowd collaborators are then not compensated despite effectively completing the HIT; this reporting provides one mechanism to eradicate that behavior.

8. *State average payout speed for HIT(s).* Simply because their labor occurs in a distributed fashion does not mean crowd collaborators are less deserving of a regular, predictable paycheck. Academic requesters should make extensive effort to review submissions within 24 hours and release payment at that time. Given the varying speeds it takes the platforms to transfer that compensation to the collaborator, this helps collaborators estimate their future earnings with better accuracy.

2. *Privacy and security.*

    1. *State technical format of task; e.g., were collaborators required to open a new browser window (distinct from the HIT page on the platform's website) or download any additional software? State all format(s) and rationale for each.* Previous work [108, 106] has demonstrated that working on crowd work platforms introduces an individual to a number of cybersecurity and privacy concerns. Where possible, the activity for a HIT should be contained

in the official HIT page on the crowd work platform. If additional screens or software are necessary, crowd collaborators should be informed why those steps are necessary and how they will appear to reduce surprise and give crowd collaborators a chance to consider whether or not they feel comfortable engaging in the HIT. There are also considerations raised by [418] about the limitations of HCI work on piecework platforms which should be considered. One of the data workers described their initial impressions of HITs on one platform: "*some of them are kind of weird*" in reference to a posting that asked the worker to upload selfies and another that asked for a copy of the worker's government ID. "*Think it's kinda sketch but I'll do it*," another worker said of tasks that required them to open new browser windows to a provided link.

2. *If collecting user demographics, was a "prefer not to answer" option for all questions available and were collaborators clearly informed that they would not be penalized for selecting that option? Further: were collected demographics protected by a privacy protocol and was this protocol made available should collaborators want to see it?* One benefit of micro task platform sites is that they often allow a crowd collaborator to remain anonymous to the requester. This may allow some individuals to gain an income where they might otherwise be unable to engage in work for fear of discrimination, persecution, or ridicule. While collecting demographic information may be important (2.4.1), care should be taken to allow individuals to protect some (or all) of their demographic information. Further, as many tools to help automate collection of HIT responses automatically record possibly identifying informa-

tion (e.g., location and/or IP address), special care should be taken to protect potentially identifying demographic information, even when the HIT activity does not require sensitive information. Individual collaborators may not want to identify as being such for any variety of reasons and careful care should be taken to prevent them from being deanonymized, even if the likelihood is extremely low. For example, [419] illustrates the need for privacy for low-income women in the Global South.

3. *HIT structure and format*.

    1. (A) *Average satisfactory completion time in trial runs;* (B) *trial population and size of population task piloted with; and* (C) *approximate relationship of population to crowd collaborators.* HITs should be tested for both functionality and estimated time to complete. Requesters may try to determine this information but are often incorrect [94]. In part, the validity of the approximation of the pilot population to the crowd collaborator population may be difficult to ascertain. Chapter 3 of [420] provides a starting point for comparing key demographic factors of crowd collaborators compared to the pilot population and can inform assumptions about approximation validity. In Drs. Mustafaraj and Lerner's labs, student researchers working on different projects pilot each other's studies; however, given the topicality of each lab and that student researchers generally have high literacy as college students who have been trained in such, additional time is added to compensate for crowd collaborators who may not have had the same opportunities or have the same general familiarity with the topic or task type.

2. (A) *Range and median of completion times for accepted crowd collaborators;* (B) *range and median of completion times for rejected submissions by crowd collaborators.* Requesters should pay attention to the amount of time required to complete their HIT. Deploying HITs in small batches provides one way to ensure that completion times and compensation for such are fair—if this ratio is not reasonable, future batches can increase compensation for the task and past collaborators should be compensated additionally via bonuses as appropriate. Requesters should also be aware of the *polychronicity*—or multitasking habits—of crowd collaborators on some platforms [417] and put the HIT upper limit at a generous time allotment.

3. *Confirm use of persistent progress bar or other indicator of progress*. The use of a progress indicator allows crowd collaborators to determine time spent on the task so far (relative to compensation) and helps them make an informed choice about whether or not to continue with the task. This should not be an issue in HITs that have calibrated payments to time spent with accuracy.

4. *Data collected*.

1. *Describe any steps taken to root out automated responses or malicious entries.* CAPTCHAs and "attention check questions" (often simple calculations, e.g., "what is two plus three?", or hidden directions, e.g., "regardless, check the fourth option below" after a long block of question text) help requesters root out automated or insincere entries [20]. However, requesters should ensure that their methods are accessible to collaborators who have hearing or visual impairments and may be using alternate technologies. Esti-

mated time to complete these authenticity/sincerity checks should be compensated and payment should consider the time it may take a collaborator using assistive technology to complete. The Dynamo guidelines also suggest double checking functionality of all attention check devices [413].

2. *Whether or not crowd collaborator demographics were collected; rationale for choice; and basis for demographic categories (if used).* There are a variety of reasons for which requesters may or may not choose to collect particular pieces of collaborator demographics. In some cases, particular demographic experiences may be correlated with cognitive biases that affect how the ensuing dataset should be understood [414]. In other cases, collecting demographic information may require extra labor from crowd collaborators, which can be frustrating when the cause for collection is not clear [413]. One of the data workers described situations in which they felt their demographic background (as it shaped their experience) was relevant, citing image recognition in a case where they felt it mattered depending on the kind of image being labeled. In contrast, if they were providing textual translation of a photographed word, they felt it was less important. In cases where demographics were requested, one data worker suggested that the requesters should share their own demographic background, to help contextualize the work and help the collaborator gauge the motivation of the request, which all the other data workers present agreed was important. If demographics are collected, the language used to request that information should be carefully considered and respectful of the diversity and variety of human experience and background. For

example, Scheuerman et al. demonstrate the reductive language used in computing around gender that does not reflect the diversity of gender in the human population at large [84]. Free-text options and multiple-selection checkboxes may facilitate this, along with an opt-out choice for all questions, as described above (2.2.2).

5.1.6 Future directions

We view the proposed cover sheet as a "living" document: we hope to accrue feedback for the contents of the cover sheet through the SIGCHI community, to present a version for practical use that covers as many concerns about just labor practices as possible. We will post these preliminary guidelines on GitHub and Google Docs (& Forms) to accrue feedback[7]. We particularly seek just practices for crowd collaborators operating outside of the United States or who are undocumented in the US, given that our experience is both US and documented-resident centric. For example, currency and payout format may be concerns we should investigate more deeply when the default on some platforms (e.g., AMT) is documented US residents operating in the US. We also hope for suggestions about accessible formatting of HITs, such as those introduced by [109] and accessible platforms such as BSpeak [110]. Finally, we seek to engage with professional digital pieceworkers to compile their feedback and will investigate respectful, collaborative ways to engage with that community.

Following a feedback cycle, we plan to conduct in situ experiments with academic researchers utilizing crowd work platforms for ML data work to see how the cover sheet affects their work, both in how they deploy their tasks and how they later use the data collected. We will then explore how to institutionalize the practice of mandatory reporting crowd collaborator employment terms in venues where such work is presented—for example, ML community conferences and gatherings.

[7]Links collected here: `https://annabelrothschild.com//projects/alt.chi-22/pro-social_crowd_collaborator_recruitment_guidelines`

We believe that the push towards automation and the ML training dataset development that requires an immediacy of action to ensuring proper behavior by academic requesters. While many academic requesters using crowd platform labor for this purpose may be interacting with crowd collaborators with sincerity and best intentions, it is still necessary to push for institutional norms that guarantee just treatment of crowd collaborators. There are also other institutions—for example, individual Institutional Review Board (IRB) programs in the United States—and funding bodies, as well as professional associations (such as the Association for Computing Machinery), who should be considered as sites of enforcement. Along with supporting high-level pushes, like that of the European Trade Union Confederation [421], we hope to provide immediate improvement in the conditions of workers on crowd labor platforms, particularly those used by academic researchers for ML data work.

5.1.7 Conclusion

In this work we propose a cover sheet to accompany the submission of projects to academic venues that require the labor of crowd collaborators. Our goal is to surface the conditions in which crowd collaborators are operating and ensure that academic requesters – specifically those seeking ML training data—treat crowd collaborators fairly and respectfully. Through alt.chi we seek feedback on the first iteration of the cover sheet and hope to discuss aspects of crowd collaboration terms which we may be overlooking, such as concerns of crowd collaborators located outside of the United States, along with those that utilize assistive technologies to work on crowd labor platforms.

5.2 Understanding Requesters' Understanding of Data Workers

I turn now to how requesters of data labor used to train or refine AI systems understand data workers. In order to put something like the cover sheet described above intro practice, it must be placed strategically so that requesters will encounter it and understand the value represented by the cover sheet towards workers.

5.2.1 Methods

To understand the requester–worker relationship, we[8] interviewed 52 participants between November 2022 and January 2023. This was a diverse pool of participants from both academic and industry settings, who acted as requesters for both research projects and commercial products. Our study was approved by our Institutional Review Board (IRB). The interviews were conducted by 87 students of Georgia Tech, along with the first author (acting as a TA), enrolled in a qualitative methods for computing course, taught by the last author. Interviews served as the basis of the course final project, wherein students were tasked with identifying participants and conducting a semi-structured interview using the interview protocol created by the authors. The objective of the research and interview protocol was thoroughly explained to the students before they began work and all students were part of the IRB's approved human subject protocol. They were encouraged to speak to a requester who has posted a task for data annotation, with each student expected to speak to one requester.

Participant selection

The criteria for participant selection was to include industry professionals and academics who have used a crowdsourcing platform (e.g., Amazon MTurk) to source data or data work for use in a large, data-intensive system (mainly for AI and ML). Potential participants were identified by the students through a range of sources like forums for particular digital task platforms and students' professional networks, along with those of the course teaching staff. Students recruited participants through word of mouth, email and social media posts. Participants were not compensated in any way for their participation in the study, due to Georgia Tech regulation.

Due to the course timeline constraints, students unable to identify a requester focused

[8]Work done collaboratively with Ding Wang, Niveditha Jayakumar, Lauren Wilcox, Carl DiSalvo and Betsy DiSalvo, which appeared at AIES in 2024.

on AI or ML tasks were allowed to interview someone who had served as a requester for a different kind of project, such as a psychology study. Therefore, we removed 35 interviews from our corpus because the interviewee did not describe or contextualize their work as data collection or work for a larger, data-intensive system, namely, for AI or ML.

Interview structure

Semi-structured interviews [309], following the interview protocol, were conducted with all the participants. This was followed by a 10 minute think-aloud session where participants walked the interviewer through the platform they had used for crowdsourced data annotation. All interviews were conducted via Zoom. On average interviews lasted about 45 minutes. At the end of the interview, participants were tasked to fill out a demographic survey (age, gender, employment status, and educational background). They also filled a separate form to express their consent (if granted) to be contacted for follow-up interview. Interviews were then transcribed and anonymized.

Interview questions focused on the participants' experience of crowdsourcing and perception of the workers who completed their task(s). Participants were also prompted to share more about certain nuances like payment calibration and metrics to qualify workers for a task, as part of the think-aloud session.

Data analysis

The first author first developed a preliminary codebook after reviewing all interviews and performing open coding [422], in consultation with the last author (course instructor), who had listened to each student group present their findings from their interviews as part of the final project. The whole research team then read a selection of interviews and updated the codebook, before the first and second authors re-coded the entire interview corpus with the new codebook. Throughout the re-coding process, incremental changes to the codebook were made in consultation with the larger research team through weekly meetings

over about three months. After the re-coding was performed, the research team discussed emerging themes and refined the open codes via axial coding, based on the grounded theory approach [307].

Participant backgrounds

We highlight four aspects of participants' backgrounds, as known to us, to help contextualize their experiences:

- *Work context*: We roughly delineate participants into "research" (38) and "commercial" (14) requesters. This distinction is based on the purpose of the tasks the requester makes; for tasks that fuel or refine a commercial product, the researcher is "commercial" (e.g., a voice technology startup). Fundamental knowledge or insight is grouped into "research" requests, e.g., those by a university professor's lab, to test the naturalness of generative text systems. From this background factor, we can make sense of where interventions might be most useful (e.g., research conference vs corporate guidelines).

- *Learning how to 'request'*: We are concerned with how requesters come to understand a platform and design tasks for it; one site of intervention might be in the materials requesters use to learn how to create a task on a platform. Several participants used multiple resources. Co-workers or collaborators were a frequent source of instruction (17), while another 15 participants turned to search engines and discovered articles, blog posts, and GitHub repositories with suggestions. The same number (15) referenced platform-published documentation. Another nine participants learned from YouTube tutorials, with eight receiving instruction from their academic advisors. A fair number (7) consulted no explicit instruction materials and instead learned from self-exploration of the platform. A much smaller number (3) received explicit support from the company operating the platform.

- *Estimating task timing*: Task timing provides some preliminary insight into how participants made sense of the workers, though not all participants described their model for estimating task timing. For example, how requesters compare their own reading speed to that of potential workers. The majority of participants (12) did a rough, self-determined estimate of how long it would take workers, while nine asked their fellow lab members to try the task and average their time taken. A small number (3) conducted a formal pilot test with workers to get an average completion time.

- *Calibrating payment*: Of those who described their payment model, the most common model (11) for determining payment was paying their local minimum wage (often $12/hour) based on how long they thought the task took to complete, with an additional two paying "slightly more" than minimum wage. Some participants determined how much to pay workers based on direct instruction from a superior (advisor or boss) (2) or as derived from total budget divided by number of tasks (3). Others (5) differentiated based on what they saw as the types of questions or tasks, to which they had a pre-assigned mental model for appropriate payment. Several participants (11) assigned a random amount of payment (e.g., 50 cents) they felt was appropriate, or assigned a random number. A small number (2) followed what the platform suggested (amount varied by platform).

Participants requested tasks in three main categories (interviewees may have multiple kinds of task requests):

- *Data annotation and classification* (30 total): Including text annotation (e.g., sentiment annotation, or rating violent intention in social media posts) and image labeling (e.g., drawing bounding boxes, or describing image contents). Generally refining or labeling an existing dataset for use in model training.

- *Data collection* (12 total): Used to procure a dataset for model training, including tasks like requesting YouTube viewing history and collecting speech samples and

text generation for NLP projects.

- *Tool or system feedback* (19 total): User testing or reviews of AI or ML systems. Sometimes of a non-standalone algorithm, such as rating accuracy of image classification system that will be incorporated into a larger product. Other examples include rating or describing experience with a chatbot, Explainable AI system, or the naturalness of a generative NLP system.

Participants used a variety of digital task platforms, with many participants active on multiple, including Amazon MTurk (46), CrowdFlower & FigureEight & Appen[9] (8), Prolific (5), Microworkers (1), Upwork (1), Fiverr (1), Dial Crowd (1), Labeler (1), Tokola (1), iMerit (1).

5.2.2 Findings

Our findings are divided into three sections: the first section details how participants, in their role as requesters, perceive the workers they hire through digital task platforms. The second section examines the proxies these participants used to identify or select workers who meet specific criteria. Third, we present the impact of these proxies on the datasets emerging from these requester-worker dyadic interactions, highlighting the data curation process, especially regarding the systematic exclusion of workers' perspective from the final dataset.

Requesters' perceptions of worker identity

We observed significant variation in the vocabulary requesters used to describe platform workers. The workers employed by our participants were referred to as "users," "participants," as well as "workers." Interestingly, individual requesters often employed several of these terms interchangeably during their interviews to describe the same group of workers.

[9]CrowdFlower was renamed FigureEight, which was then acquired by Appen in the span of a few years; we report these in a group since some participants referred to the company by its name at the time of requesting and others by its current name.

The term "employee" was notably used just once, by P42, to draw a strict distinction between platform workers and the full-time annotators (referred to as "employees") at P42's company.

Who requesters think workers are There are two prevalent paradigms for hiring or determining worker eligibility for tasks on digital task platforms. On non-anonymous platforms like Upwork, where workers have a clear identity visible to the requester, selection hinges on the worker's portfolio and identity. In contrast, on pseudo-anonymous platforms, such as Amazon MTurk, requesters cannot be certain of the worker's identities. However, they can either apply platform sponsored filters, or create their own screening tasks to target specific groups, like college students. This subsection focuses on the latter model, specifically the requester-led screening processes typical of most pseudo-anonymous digital task platforms, including, but not limited to, Amazon MTurk.

Workers on digital task platforms can theoretically be anyone with an internet connection and a suitable computer, subject only to platform-specific requirements regarding nationality, residence, and language skills. While the exact demographic and experiential composition of these workers has been estimated in various studies over the years, we found that requesters hold only a vague understanding of the potential worker pool. This observation aligns with what Kapania et al. noted in their 2023 study [92]. Many requesters perceived the worker base of digital task platforms as "the general public". For instance, P31 explained that they imposed no specific restrictions on which workers could work on their tasks, aiming for *"the general workers."* Similarly, P36 admitted to having *"no idea"* who was completing their tasks, describing the workers as *"random people, random Turkers."* P6 echoed this sentiment, stating that for their latest task, *"the background of the participants were not identified as a crucial factors for the result—we were just strangers on the platform."*

A minority of requesters had specific, albeit often anecdotal, insights into their worker

population. For example, P1 who had built long term relationships with a handful of Amazon MTurk workers over months, learned through regular communication that some were working from home or supplementing their government-sponsored disability benefits. Consequently, P1 formed a general impression that many their workers were similarly working from home. However, P1 acknowledged this view was based on those who communicated frequently, noting *"[I] mostly imagine[d] those people even though I know they're not the majority of people...they're just the ones who communicate more."*

At the same time, requesters aiming to hire from the general public, members of which they presumed to be accessible via these digital task platforms, implemented various measures to secure "good data." A frequent concern raised by our participants was the quality of the data (or task performance), as it contributes to the larger outcome dataset and subsequent model development. While good data was highly valued, the biggest threat to it, in the perspective of our interviewees, was the presence of "bad actors".

Good data vs bad actors Although our participants had only vague notions of who was completing their tasks, they expressed significant concerns about "bad actors"—workers who submitted poor or fraudulent responses just to receive payments. Notably, while the data quality produced by a worker might be deemed "good," individual workers were never described in any of the interviews as "good". This implies that requesters, content with the data quality, seldom question the identity of the worker. For example, P5 believed that offering a minimum of $20 per hour would ensure *"good annotations"*. This reflects a common belief in freelancing that higher pay attracts better candidates. However, P5's focus was solely on the data quality, which is separate from the characteristics of the person producing it.

In contrast, when dissatisfied with the received data, many interviewees expressed frustrations towards "bad workers" whom they saw as performing the tasks poorly, sometimes even with malice. In these cases, the requesters attributed the poor data quality to the

worker's shortcomings. For instance, P9 spoke about implementing quality checks to avoid workers who *"usually cheat in these platforms."* P39 explicitly described methods for *"filtering out bad actors,"* implying that such actors were responsible for undesirable data. Similarly, P13, who generally accepted most submissions, stated that their only reason for rejection was when *"someone deliberately tries to poison the data."* P13 further explained their methods for manually reviewing task submissions to determine which workers *"are doing bad."*

Many requesters from an academic setting, who posted tasks as part of studies, saw workers who completed tasks primarily for financial compensation as undermining the integrity of their research. These academics expected participation driven by a sincere interest in contributing to the development of new knowledge. Even though P44 recognized that *"Amazon Turkers is a group of people that does this for money,"* they were frustrated by workers or research participants, who they felt were *"just clicking, you know, like randomly or very quickly, to get you [the requester] the [completed] survey to get money."* This is only one telling example of the disconnect between the academic expectation of earnest participation and the practical motivation of workers on these platforms.

The *good data vs bad actors* dichotomy reflects a broader dynamic in the relationship between requesters and their workers. When task submissions, contributing to the larger dataset, meet expectations, requesters typically don't question the workers' identities, expertise, or experience. Yet, when the data is presumably unsatisfactory, the focus shifts to the workers as individuals, holding them responsible, as a *bad actor*, for the subpar outcomes. This shift underscores how workers' human identity and experience are overlooked when things go well but scrutinized when issues arise.

Proxies used to select workers and determine submission quality

To discern workers who would produce high-quality data from potential bad actors, our interviewees—data requesters—developed a number of *proxies*, or micro-tests to approx-

imate whether a worker delivers a certain quality. We categorize these proxies into two types: those determing a worker's identity and suitability for a particular task, and those evaluating their ability to produce high-quality (or desirable) data. Notably, many requesters used multiple proxies in tandem, filtering out elements of what they believed to be the general (implied US) population. For example, P11, who was interested in recruiting members of the general public to their task, expressed a desire to avoid biases such as English fluency in their task outcomes. However, P11 also put in a number of common checks, including filtering potential workers based on their prior task acceptance rate.

Proxying worker identity The following is a synopsis of the proxies used to establish the identity of the workers, on platforms that operate on a quasi-anonymous model.

- *English fluency*: Interviewees, all based in the United States, frequently emphasized the importance of potential workers' English fluency. P30, working on data collection for NLP tasks, prioritized hiring *"English native speakers"* and used the platform's settings to filter workers accordingly. P19, on the other hand, described their strategy for ensuring English fluency by selecting workers from locations where English is a primary language. Though they admitted to not finding a reliable method on Mturk *"to ensure that that requirement is met."*

- *Age*: Due to legal requirements, many interviewees stipulated that workers must be at least 18 years old. In academic research tasks, requesters typically asked workers to simply affirm they were were 18 or older, aligning with standard practices for U.S.-based human subjects research. Conversely, in commercial or industrial research settings, interviewees more often relied on platforms' official filters and purview for age verification.

- *Location and time zone*: Using location based proxies was popular among requesters doing geographically-specific projects. For example, P16 wanted workers who were

familiar with U.S. news and events and limited their selection to U.S. residents only. Similarly, P18, who was interested in having data annotated with various English dialects (UK, Australia, Canada, Ireland, New Zealand), filtered for workers from these specific countries using platform features. However, they were cognizant of workers using VPNs to appear as if they were from these locations. P37, who also sought workers from the UK, Australia, and the U.S., posted tasks during regular working hours in these regions, instead of the requester's own local timezone. Additionally, they informally reviewed long-form text submissions believing *"from the way they [the workers] speak,"* they could identify the worker's *"ethnic background"* and whether it aligned with their targeted regions or not.

Proxying aptitude These proxies were used to ascertain how well a given worker would complete—or did complete—a given task. We group these proxies into pre and post hoc, or those applied to test worker aptitude based on their platform-specific track record or those used to test how well a worker performed on the requester's task.

Pre hoc proxies:

- *Approval rating*: Requesters commonly used approval ratings, which reflect a worker's work history and aggregate ratings from tasks completed for other requesters, as a filtering criterion. However, what constituted as an acceptable prior approval rating varied between requesters: P47 considered a rating above 95% as *"good"*, while P10 set their minimum at 98%, and for P11, the threshold was even higher at 99%.
- *Number of tasks completed*: Requesters often gauged worker's reputation by their previous task history, i.e., the sheer number of tasks they completed. Similar to approval rating, the threshold for desirable task history varied greatly. For P21 that number was 1,000 prior tasks, where P48 set a benchmark at 100, and P10 at merely 50.

Post hoc proxies:

- *Keyboard interaction*: How much a worker engaged with the keyboard—and the speed of keyboard interactions—were of interest to requesters in trying to ascertain how sincerely workers completed their tasks. P16 wrote a tool to check if workers copied and pasted into free-text boxes, in order to ensure that data collected *"was not useless."* P12 assumed that a fast repetition of clicks meant the worker was a bot, and so employed a keylogger to make sure workers were genuine with their replies; P12 wanted also to avoid workers who were just *"trying to get their two cents or whatever."* P39 performed an *"entropy analysis"* of the amount of time spent on the task, premised on the idea that if the worker spent far less time than the team determined appropriate in trials, the worker's submission would be rejected.

- *Answer pattern*: Requesters saw the visual pattern of answers or choices (e.g., on sequential multiple choice or Likert scale questions) that workers selected as being a signal of genuine effort. For example, if a worker always selected Option B in a multiple choice question, that was a red flag for P12. Similarly, P3 would manually review all task submissions, checking if participants choose only the first option, or another pattern.

- *Attention check*: Requesters used assorted wording (e.g., *"quality checks")* [P9], *"dummy questions"* [P52]) for what is commonly known as an *attention check*, or verifying a worker is both human and focused on the task at hand. Requesters commonly complained that verifying attention check results was time-consuming, and perhaps *"the most time-consuming activity"* of posting a task (P9). P34 explained, echoing others, that they began including attention checks because they *"didn't trust them [the workers] anymore."* CAPTCHAs are commonly used to filter out bots, while reading comprehension questions are often used to make sure workers are carefully reading directions. For example, P1 tried to write what they felt were fair attention checks, or ones that are *"really, really obvious where you can only get them*

wrong if you're literally not reading...it's a multiple choice question that says do not choose the other option, the other options says you have to choose this option."

- *Gut reaction*: Several requesters relied on a *gut reaction*, or instinctual feeling of whether or not task submissions were correctly done. P36 had a two-part strategy for verifying quality of worker submissions, the second step being *"more or less ad hoc...[it] was subjective...because we would look at the annotations and be like, 'okay, yeah, this person has been producing consistently good annotations, so he should continue with it."'* P50, meanwhile, described going through submissions for a generative task manually and deciding *"these ones are good, these ones are bad,"* a process they ultimately found *"annoying."* Perhaps the best summation of the nature of attention checks is the experience of P21, who used a series of attention check methods and in turn received feedback from workers in which they called P21 an *"evil genius"* due to the difficulty of properly completing the task.

- *Coherence as trust*: Comparing a worker's answer or selection for a subset of task activities was a common way to determine how correct a submission was. Two methods were used to do this:

  - Majority consensus: Picking the most commonly submitted label or annotation and checking all submissions against that correct answer was a common process to automate accepting and rejecting submissions. P2 set a threshold for agreement and once that agreement was met, accepted that label as the correct one. P16 similarly employed agreement scores between workers to determine the acceptable annotation.

  - Gold standard adherence: P5, for example, checks the agreement score of different annotators against their (requester developed) answers and if the score is high enough, accepts the submission. Similarly, P50 used a small pool of self-labeled data to test worker submissions against.

One other indicator of potential worker submission quality used by requesters operating on MTurk was the Amazon-designed *Master Turker* distinction. Of the qualification, Amazon vaguely states that Master Turkers have completed thousands of tasks on the platform and have done so with a high level of performance.[10] Notably, hiring only Master Turkers as a requester on MTurk carries an additional platform-imposed surcharge. Requesters in our corpus had mixed feelings about Master Turkers and the meaning of the designation. Some felt it was an assurance of reputable work; P51 posted tasks only for Master Turkers, feeling their work quality was far superior to those of non-Masters. P20 felt the surcharge for Master Turkers was worth it, since *"we ran a cost analysis and realized that it was actually cheaper to pay significantly more for Master Turkers because the results that they gave us back were just so much better."*

Others, however, felt the designation to be relatively meaningless. P37, for example, stopped using Master Turkers after reading on a forum that the same level of worker could be selected for using a mix of other filters, including prior approval rating. P18 felt that using only Master Turkers was contrary to their goal of capturing a broad cross-section of workers to get annotations beyond those that would be generated in an exclusively academic research environment, since they felt Master Turkers represented an elite within the larger MTurk worker population. Similar to P18, P5 was not sure the Master Turker designation was meaningful—they were also concerned with workers being unsure of how to obtain the designation, which they saw as unfair.

Curated datasets

Reflecting on the larger patterns of data curation described in these findings, there is a broader perspective in their combined presence. To synthesize the findings presented thus far, requesters posting ML and AI tasks on these semi-anonymous digital task platforms attempt to ascertain the following:

[10] https://aws.amazon.com/blogs/aws/amazon-mechanical-turk-master-workers/

- what they consider to be task-pertinent information on a worker's identity
- a promise of quality work based on past performance
- attempts to confirm the quality (sometimes authenticity or sincerity, depending on task type) of worker submissions.

Requesters' conception of workers generally follows one of three models. In the first, requesters simply don't care who the workers are as long as the task is satisfactorily completed, leaving open questions about the ability of these datasets to generalize over diverse populations. Second, requesters believe their workers represent the *general public* or a specific subgroup within that public, but don't actually know this to be true—instead, they refer to a handful of now-aged studies that suggest this might be the case. Third, requesters see workers as research subjects or study participants, and treat them as volunteers, rather than a compensated workforce. This demonstrates a fundamental misunderstanding between workers and requesters, one that has to be reconciled before either party will be happy with their shared interaction. However, more broadly, these models present a problem for the datasets produced, in that these datasets are mistakenly generalized to a population (unknown, general public, or study participants) that are not a reasonable descriptor for the worker population who produced them.

We now return to the idea of proxies—or measures of things that requesters use and believe to be reasonable approximations of the true value they want—which present problems for requesters and workers, both. For requesters, developing and implementing these proxies is difficult; while several participants mentioned reading online forums (e.g., P37) or using metrics they saw in academic papers (e.g., P28), there is no standard, uniform way to ensure that a proxy has the right threshold. As much as requesters struggle to develop proxies, workers must also re-qualify for tasks they may already be skilled in, wasting their time and energy.

Further, we question whether these proxies were accurate approximations of the value

that requesters hope they were. Consider, for example, the emphasis on *"native"* English speakers, as opposed to being fluent in English. Why could an immigrant to a majority English-speaking country be unable to speak English sufficiently enough for the task? Given that the types of tasks requesters in this interview corpus are posting are generally with large-scale ML or AI applications, why would you want to exclude the variance of vernacular English?

Proxies are a burden for workers as well; they must repeatedly demonstrate their skills for every requester they encounter, hoping to meet the task-specific discretionary threshold. Proxies may also be a poor test of actual domain experience and knowledge. Consider the minimum time limits many requesters use to track sincerity of worker effort: if a worker spends much of their day (given that many digital task platform workers do so as their main income source [135]), for example, drawing bounding boxes for image labeling. They are, given a new task that requires that same skill, highly likely to complete it far quicker than a requester who has maybe done only a few dozen for practice. Workers must also spend time qualifying for tasks through pre-screeners—while some requesters still compensated workers for these pre-task activities, many more did not.

These proxies also force workers to engage in sub-standard work should their ratings fall due to simple misunderstandings, or even malicious requester actions. The Turkopticon "End the harm of mass rejections" campaign, for example, highlights how requesters can take advantage of workers, basically by collecting their labor and failing to pay them for it, forcing workers to pick up ultra-low paying tasks in order to rebuild their ratings [125].

In adopting these proxies, requesters edit out what they see as undesirable submissions, with the goal of curating a dataset of what they see as quality data points. We observe contradictions in this process; namely, requesters who seek input or submissions from the *general public*, but then edit out variance within that larger sample, due to the proxies they employ to obtain and verify worker identity and quality.

5.2.3 Discussion

In this chapter, we highlight the issues encountered in the current practices of perceiving and treating workers on digital task platforms. Rather than direct assessment of their individual skills, workers are evaluated against vague proxies designed for behavioral assessment at scale. These practices pose challenges not only for the workers but also affect the integrity of the resulting datasets. In turn, these datasets, often skewed by the underrepresentation of minority voices, form the foundation for broader AI and ML systems, impacting both the general public and the developers of these systems. In the following subsections we propose strategies to safeguard worker rights and critically examine the implications of sourcing datasets through digital task platforms, especially in the context of representing the general public.

The problem with proxies

Returning to the impact of quantification in platform-mediated data work, as discussed by Espeland et al. [423], proxies emerge as a prime example. At a fundamental level, proxies present a concerning practice: as validation implements, they are of questionable *construct reliability* and *construct validity*.

Drawing on Jacobs & Wallach's exploration of fairness and bias, *construct reliability* is akin to replicability [424]. Applied to proxies, there is little replicability among requesters, as each implements a conceptual proxy differently. Within our corpus, there were no universal practices for any given proxy. Instead, requesters often rely on general knowledge from professional or academic sources to create their own interpretations, leading to significant variance in criteria like "significant prior work history" as measured by completed tasks. This raises a critical question for workers with an established, reputable profile on one platform: how can they transfer these credentials to another digital task platform? Currently, there are no mechanisms to facilitate the transfer of a worker's CV or portfolio between platforms, posing a challenge for worker mobility and recognition of individual

workers' skills across different platforms.

Proxies not only fail *construct reliability* but also lack robust *construct validity*. According to Jacobs & Wallach, *construct validity* entails ensuring that measurements are accurately grounded in the intended construct and encompass all its relevant aspects [424]. The proxies requesters used do not meet validity in this regard, as there is no consensus or evidence confirming these measures accurately and comprehensively represent worker's identity or the quality of task completion. These proxies seldom offer a fair assessment of labor. For instance, some proxies arbitrarily assume priority of *native* speakers over *fluent* speakers, while others aim to achieve a general consensus, thereby minimizing diverse perspectives. This approach can lead to the silencing and even penalization of workers who express minority opinions or unique lived experiences, as noted by Kapania et al. [92]. Further, domain experience is paradoxically undervalued, as seen when workers who complete tasks rapidly are excluded from the final dataset. This raises significant concerns about the fairness and efficacy of proxies in accurately assessing and valuing the contributions of platform workers.

Proxies not only symbolize the lack of trust requesters have in the workers but also lack both construct reliability and validity. Additionally, they impose unfair expectations on workers. Requesters frequently use attention checks and review visual answer patterns, expecting a level of focus from workers that is often unattainable or simply inaccurate. This raises an important question: Why is there an expectation for digital task platform workers, many of whom rely on these platforms as their primary income source, to maintain intense concentration for their entire workday or on each micro-task? It's unlikely that these requesters themselves maintain such hyperfocus throughout their own multi-hour workdays. This disparity highlights a potential double standard: Why is it deemed fair to expect digital task platform workers to exhibit a level of focus that is not expected in other professional contexts? This contrast demonstrates, once again, the power imbalance between data work annotators and requesters [15, 91].

Labor protections for invisible collar workers

Requesters expressed concerns about workers' motivations on digital platforms, suspecting them to be primarily financially driven. However, it's crucial to recognize that these individuals, operating as *workers*, face limited job security and scant prospects for advancement with any particular requester. Their primary interests may not extend beyond meeting the minimum quality requirements. This contrasts sharply with the paradigm in citizen science, where volunteers undertake data tasks mostly for personal fulfillment or a belief in a larger scientific or social mission [425]. This sense of purpose and ownership that citizen science volunteers have towards their work likely translates into greater investments in the quality of data.

The prevalent use of proxies on digital task platforms highlights the need to fundamentally rethink the requester-worker relationship, with a focus on recognizing the realities faced by both parties, especially the challenges that workers encounter due to a lack of workplace protections. As discussed earlier in this chapter, the pseudo-anonymous digital task platform workers are part of an *invisible collar* class of workers. Despite earning regular income, they are not seen as employees by either the platform or the requesters. Considering the origin of Amazon *MTurk* and the vague image of a worker described by requesters, the obscured human presence of these workers exacerbates the invisibility of their labor, their identity, and critically, their rights and protections.

The recognition of digital task platform workers as employees is a crucial step in determining how to extend labor protections to them. White collar workers have compliance mechanisms, such as HR departments, which are technically meant to mediate fair treatment between employer and employee. Blue collar workers have some material protections in place, owing to the physical nature of work, such as OSHA, and the option to organize through labor unions. Pink collar work, which is perhaps the closet of the three to invisible collar labor, employs some combination of the above, given that the roles within pink collar labor may be clerical (e.g., executive assistant), physical (e.g., nurse), or a mix (e.g.,

grade school teacher) [426, 427]. While neither white nor blue-collar labor protections are perfect—union activity can face corporate resistance and retaliation (e.g., Amazon vs unions during COVID-19 [428]), and HR might prioritize legal compliance over employee well-being—they still offer more than the current status quo, which provides no protections to invisible collar workers.

Consider the recent initiative in the United States, the widely acclaimed White House Blueprint for an AI Bill of Rights [429], intended to herald in a new era of oversight for AI and ML systems. However, notably absent from this document is any mention of the people who generate and annotate the datasets that power these systems. This oversight is concerning, especially in light of the well-documented risks associated with this type of work. For example, the adverse experiences of a ChatGPT data worker in Kenya, who reported the work *"destroyed [him] completely,"* [133] highlighting the physical and emotional toll presented by this work. Additionally, there are pressing legal concerns, as seen in the ongoing cases in Kenyan and Ethiopian courts involving Meta and subcontractor Sama [430, 130]. As consumers are increasingly recognized as needing protections in the realm of AI and ML systems (evidenced by proposals like the White House AI Bill of Rights and the EU AI Act), similar attention and safeguards should be extended to the workers on digital task platforms who are are essential part of the AI ecosystem.

Rethinking dataset sourcing policies at scale

Similar to variations in proxies used, requesters are also subject to different pressures. Many of our interviewees came from academia, in which individual requesters are answerable primarily to research publication sites (such as journals and conferences), as well as the regulations imposed by funding bodies, such as the U.S. National Science Foundation. In contrast, their peers in industry may also publish in research venues, but are more immediately subject to corporate or organizational strictures. Following in the triangular understanding of the power systems presented by digital task platforms, requesters are the

ones with the ability to effect the most change as it impacts workers and, later, datasets themselves.

5.2.4 Conclusion

AI ethics necessitates better documentation and contextualization of the datasets that power AI and ML systems. While these practices have been the provenance of dataset designers, such as data scientists and system developers, we call for a turn to the the expertise of the data workers with the most intimate knowledge of a dataset's production. To this end, we conducted interviews with 52 requesters of digital task platform data work, who used the tasks to support development and refinement of AI and ML systems. Focusing on their understanding of workers whose labor made those datasets possible, we clarified requester perception of workers' identities, aptitudes, and qualifications. We found that workers are hidden from view until there is a problem; further, workers are tested with a series of pre- and post-task *proxies* meant to ensure workers are qualified to perform a task and did so correctly and genuinely. These proxies, however, are rarely formalized on more than anecdotal evidence, to the detriment of workers. Proxies, along with obscuring worker experience and perspective, contribute to a larger pattern of dataset curation, which we must understand as part of a robust program of dataset documentation. We subsequently call for a rethinking of dataset sourcing policies in both academic and industrial practice, to the benefit of both data workers and AI ethics initiatives.

5.3 Chapter Conclusion

This chapter reported on the need for pro-social interventions in data work, namely that which occurs on quasi-anonymous digital platforms. These interventions will lay the grounds for how data workers on these platforms can be engaged with as dataset auditors. This transition requires the development of genuine trust between both parties towards one another.

To figure out how the lessons from DataWorks, in how this role can be developed and shaped, to quasi-anonymous platforms, I have shared two works in this chapter:

The first section in this chapter detailed a cover sheet, to be used by academic requesters to detail—and promulgate reflection on—the terms of labor they created for the crowdworkers they employed. In the second section I dig deeper into how requesters actually design tasks for workers, insofar as they make assumptions and/or try to verify who the worker is and what they are capable of.

This goes against the notion of a quasi-anonymous platform, in which the requester is not supposed to know anything about the worker. However, requesters still find they need certain kinds of information so they can assume that the worker has done a sufficiently good job on the task(s). I find that requesters employ a series of non-scientifically founded *proxies* to gauge worker identity and ability, many of which are invisible to workers and they are unable to seek recourse or explain a situation accordingly. Since integrating broader perspective, as a matter of both error-catching and on-the-ground reflection, requires communication of those observations, what can or cannot be communicated between requesters and workers is an immediate concern.

In the discussion chapter that follows this one, I explore how these transitions can be made, in keeping with the cosmolocal model, so as not to disrupt those things about the platform infrastructure that are beneficial to both workers and requesters.

CHAPTER 6
DISCUSSION

6.1 What data workers equipped with critical data literacy and a democratic workplace setting can do for datasets

In the introduction to this work, I proposed a series of projects that would demonstrate the role that critical data literacy plays in up-skilling data work professionals to, when combined with conducive workplaces, enable data workers to act as dataset auditors. The work is informed immediately by the concepts cohered in *data feminism* [47] (see 1.3.2). It is implausible to try to make sense of the vast, involved supply chain that, and the individuals who work it, that comprise the data annotation pipeline for AI, without considering how those players interact with one another. In this dissertation, I have explored in detail how what the role of individual power and discretion is for the data annotation work that results from any given data work site. I introduce the idea of *data-worker-as-dataset-auditor* as a means of both promoting the role and expertise of data workers, which is undervalued at present [51, 88], and supports the development of safer AI system development. Safer AI systems are those based off of well-contextualized datasets, and data workers are uniquely positioned to draw on their lived experience and perspective to provide feedback on dataset contents and appropriateness. Fundamentally, I believe that data workers' lived experience and perspective, as well as professional purview, is not only valuable, but essential to the development of safer AI systems.

This dissertation began with a study of the status quo. Making sense of the an alternative to the current extractive paradigm of data annotation for AI systems [51] is to attempt to build it, which begins in the first content chapter. Here I introduce the genesis for this work: the unique setting the DataWorks organization as a field site, where the mix of

workplace-training environment (emphasizing learning over commercial returns), engaged-research vehicle (enabling open-ended workplace ethnography and novel workflows), and data services provider (as a testbed for new data work paradigms), enable an alternative to the status quo in commercial, industrial data labor providers.

But how did data work—and the data science and AI practice that necessitates it—take the marginalized labor form that it does today? First, I trace the practice of AI as a more contemporary inscription of data science, and before that, cybernetics. The same impulses of militaristic command-and-control that defined cybernetics continue to metastasize through AI. To imagine alternatives to this status quo, and to proceed with contesting (via generative critique) how those alternatives can be built, takes us one step beyond the data itself. This is where I turn to *cyberfeminism* (see 1.3.1) as a way to understand how an alternative practice to the current, cybernetic-influenced order came to be.

Following the methodological flexibility and embrace of standpoints, complexity, and epistemological chaos, the rest of this dissertation is formatted on the conceptual basis and agnosticism that shapes cyberfeminism. Notably, in order to conceptualize and address data work as an ecosystem with multiple stakeholders and parties, including different groups of professionals, constraints and opportunities provided in the form of various software and hardware, along with larger, interpersonal and community-level social dynamics, there are a diverse set of approaches, methods, and interventions included in this document. No one method or product would be appropriate to fix the varied but indivisible set of problems that this dissertation makes a starting point towards addressing. A singular technical product cannot, for example, reset an entire group social dynamic.

The first chapter also contains a parable for an alternative way of building and utilizing datasets: turning to the practices of civic and non-profit data workers. These professionals find ways, both ordinary and inventive, to re-contextualize the datasets they work with. These methods range from established collaborations between, for example, municipal departments, to informal collaboration between individual professionals working with similar

datasets. This work inspired much of the later shape of DataWorks as an organization, as well as drew my attention to workers and data practices that are in contrast to the contemporary data procurement paradigm for commercial AI systems.

Through DataWorks, and with the invaluable collaboration with several generations of Data Wranglers and Data Fellows, I demonstrate that in the context of this unique work environment, critical data literacy enables an exciting turn of events. This is the proposition of the third content chapter. Unlike the status quo in data labor, where data workers tend to know nothing about the project nor its objectives, for which they are labeling, standardizing, or cleaning data, critical data literacy gives data workers the technical skills to employ their existing lived experience and perspective to engage with clients to make sense of the role of the workers' labor, with resulting implications for the shape and context of the dataset. Those impacts are not abstract; in the case study of the AAVE project, I am able to share the concrete range of changes, both material (structure of data annotation task) and conceptual (in the design and deployment of the resulting system and work, particularly in its social, not only linguistic, contextualization of AAVE). In other words, what results is a dataset and a generative NLP application that is not only socially sound, but is more respectful of the origins of the linguistic phenomenon under examination.

The AAVE case study is coupled with the Datum Fieldnotes tool, as a way to concretize and expedite the kind of dataset inquiry and auditing that took place in the case study. Namely, while the dataset auditing that took place in the case study was ad-hoc and unplanned, emerging naturally from data workers' concerns, it should not have to be this way. Instead, given what we now know about the potential of critical data literacy to help data workers translate their observations, perspectives, and feedback on a dataset or data task, particularly with an eye towards such a process happening in a larger or more rigid organization, we need tools to streamline and formalize such inquiry. Further, the very presence of such tools in the data work office setting suggests that the act of dataset inquiry and review is a natural part of the data labor process. This is counter to the sta-

tus quo in platform-mediated data work, particularly that which happens at the commercial level. Instead, data workers are told to produce labor that embodies the designs of the client requesting that labor [22, 149, 51, 431].

But, where can we go from here? If DataWorks is unique, what does this change for the status quo, for the majority of data workers globally? Is auditing possible at scale? This is where I return to the idea of the cosmolocal 1.3.3.

6.2 What does it mean to scale-up in the world of data work sites?

Before continuing to explore the second half of work that constitutes this dissertation, I return to an important theory of scale: that of *cosmolocal* [159] (see 1.3.3). The notion of the cosmolocal requires us to review what we want from scaling up. Specifically, what does scaling up the insights of DataWorks look like? Who is affected? My goal is actually not to mold every data labor site (including the commercial platform-mediated ones such as Amazon MTurk) into DataWorks. As an organization, DataWorks only functions because it *is* highly localized. How we identify that locale, however, is flexible.

There are multiple candidates to pick from. The premise of the University (namely Georgia Tech) provides equal part affordances and challenges: the research environment allows for certain variations from the for-profit model, as well as a the flexible nature of the organization. Because we work on the shared University campus our interactions—that of myself and the rest of the research team and the Data Wranglers and Fellows—is inherently shaped by that geographical proximity. We don't share an office, but we are physically proximate and are generally subject to the same shared University governance. What that means for the data work that comes out of DataWorks is ambiguous; I argue, however, that it enhances a sense of comradery and collaboration that is unique, though not unlike, what might happen at a small organization outside any larger institution. The projects that come into DataWorks from the research team's colleagues, as peer academics, are likely uniquely available in part due to this University site, for reasons of practicality,

belief in organizational mission, and also a need to interact with verified individuals (aka, avoid potential interactions with bots).

That locale could be the City of Atlanta; all the Data Wranglers and Fellows are from the metro Atlanta area, and the work that they perform often calls upon that lived expertise directly—for example, correcting street names and identifying locations in context would not be possible without it. In tandem with the Data Wranglers and Fellows, I too am from Atlanta—and out interactions are no doubt shaped by shared experience based out of geography, as mundane as frustrations with traffic, and as complex as the way we understand the politics that shape the datasets that clients bring us. The data that is handled and shaped at DataWorks, particularly that relates directly to Atlanta (for example, when the client is an Atlanta-area non-profit), is highly contextualized both out of professional and personal knowledge of the Data Wranglers and Fellows.

Another candidate for locale is the building that DataWorks was housed in from its inception to late 2024, or the Georgia Tech College of Computing Building (CCB). The CCB is an eponymous brick monolith—it is home to many of the instructional classrooms and administrative offices that provide the physical infrastructure of the College. To be embedded in such a place, where the anxiety of computing students around final exams is palpable, and the hallway white boards display remnants of the prior hour's TA office hours, means that one is continually confronted with the designers and developers of the very systems for which the Data Wranglers and Fellows are assembling dataset and refining the outputs of data-intensive systems. Working in this building, it is impossible to separate the downstream futures of a dataset from the potential applications.

This digression into the flexibility of identifiable locale suggests delivers a key insight into the nature of scaling up. No two places are the same, and no two data services providers should look identical. Why DataWorks *works* is because it is tied to place, a place that can be interpreted in many ways. These varied locales draw out the intricacies of locale; what is it that uniquely identified a data work site? Is it the kind of workplace? The

geography of workplace? Or the physical infrastructure in which that workplace exists? Scaling up cannot thus be easily translated into a formula for success. Instead, what is needed is a change in approach to our understanding of data work and of the professionals who practice it. The second half of this dissertation, which focuses on how these learnings might be applied elsewhere, does not suggest a transformation in physical form of these sites. This is why I turn to the cosmolocal as a way of thinking about what it means to transfer these learnings; what is needed in the profession of data work is respect for the data workers themselves, which will look different in each and every locale.

Respect for skilled labor and practitioners of it, however, is a value that we can pull from the success of DataWorks and scale-up to sites farther afield. What demonstrates respect at DataWorks might include opportunities for professional growth and development—that may or may not be the case somewhere like Amazon MTurk, where individual workers may be more concerned with the material finances of their profession, or may need safeguards (for example, against exploitative requesters) that are already present in DataWorks. It is only in a robust demonstration of respect, localized to the particularly group of data workers, and a desire for genuine collaboration that AI system designers and developers are going to be able to benefit from the insight of data workers as dataset auditors. Data workers, as skilled professionals, have both the purview (see a variety constituent elements in the larger dataset) and expertise (across projects, from lived experience and perspective) to be collaborators in the development of datasets that can be used to build and refine safe AI systems. In fact, I argue, they are uniquely situated to do so.

In scaling up from DataWorks to other sites of data labor and dataset production, following the cosmolocal model (see 1.3.3) directs our attention to first identifying and appreciating the uniqueness of each locale. Only then can we determine and subsequently carry out the mechanisms needed to develop and demonstrate the respect for data works necessary to build the conditions for dataset annotation at the pre-model training stage. Scaling up should not be frictionless; it should be messy, complicated, and entangled with the peo-

ple and social context in which the operation occurs. While there may be commonalities between the implementation of respect at different sites—for example, better pay is likely to arise, given the status quo for data work compensation broadly—the actual format and deployment will look different, pursuant to the uniqueness of each site.

6.3 Scaling up: the cosmolocal approach

In the second half of this dissertation, I shared the designs for a series of projects that would take as inspiration the work that civic and non-profit data professionals do to make sense of and contextualize the (imperfect) datasets they work with, characterized by the notion of *careful data tinkering*, or non-profit and advocacy data work shaped by care and concern for the affected populations [432].

These designs include include contributing to the work on power differentials between data annotators and the requesters of their labor. Specifically, I study the role of requesters, as the most powerful group in the triadic dynamic that occurs between requesters, platforms, and workers [416]. In pivoting the behavioral focus to requesters, I see them as the party with the greatest ability to achieve change quickly, as well as potentially having the impetus to do so. That impetus could be, as we explored in the work in Section 5.2 the reaction to what is currently an unintentional lack of awareness, or to more imposing external structures and strictures, e.g., requirements by funding bodies.

What motivates these requesters will ultimately likely be different than what it took for DataWorks clients to understand the Data Fellows' conceptual and applied skills differently. That is why the work described in Section 5.2 illustrates the gaps between shared understandings and experiences between workers and requesters; specifically, what requesters think of workers, and how well that maps to the reality of the labor experiences that workers undergo. This is conducted in keeping with the cosmolocal approach, where the emphasis is not only trying to directly to create a blueprint of the worker-client relationship at DataWorks and merely inscribing that into the known triangular relationship that already

exists.

Practicality aside (there are three key parties in the platform-mediated model, not two) doing so would remove those features of platforms like Amazon MTurk that are popular because of their form. While minor negotiations to form, e.g., adding capacity for the dataset auditor role in addition to data annotator, do not disrupt the model's core functions, requiring a more direct interface actually would. The quasi-anonymous model has many valid critiques, including that it enables malicious behavior on the part of all parties. However, it does enable a new class of data labor, in contrast to prior office-based options. Namely, workers who carry identities that might face discrimination in traditional office-based employment spaces are able to work from a private space. Similarly, given the presence of sufficient infrastructure, workers with physical disabilities or other home-based needs (e.g., caring for dependents) are able to still engage.

There are a host of problems equally associated with these affordances, however the reality is that there exists, for example, a large Turker population that works on the platform as a main income source. It would not only be disingenuous, but equally disruptive to these individuals to preclude their labor by revolutionizing the platform. Following the cosmolocal approach, the goal is not to fundamentally overhaul the shape (i.e., mechanisms for connectivity) of data labor as it appears in the platform-mediated genre, but to carefully apply the knowledge demonstrated at DataWorks with respect to the shape of the platform in question.

To this end, I developed the cover sheet project as an implementable way for academic requesters of workers on digital pieceworking platforms to understand how the minutiae of their task requests. Further, I address a gap in the literature regarding how requesters actually understand and perceive their workers. To figure out how to improve the way that requesters engage and employ workers—particularly regarding how these relationships might be reset to enable active collaboration with data workers as dataset auditors—we have to know how requesters are actually making sense of workers.

6.4 What is the next step for data workers' critical data literacy in data work?

The work described in this dissertation presents an alternative for changing way data work is currently performed and the labor conditions for those who perform it. Specifically, I offer a more pro-worker experience that produces datasets that are both better constructed and contextualized. The datasets that emerge are "safe," or actually represent the values and subjects intended by their designers. Safe datasets are the first step towards AI systems that behave as intended, because the matter (data) used to train predictive systems actually addresses the question or subjects the system is designed to consider. There are two avenues of future work that consider, in turn, 1) how to transfer these findings beyond DataWorks to other data labor production sites, and, 2) how the role of a data worker is changing in the advent of "data scarcity" and a push to automate routine data work with assistance from AI systems.

6.4.1 Part 1: Applying the cosmolocal

How can the knowledge from DataWorks about how to build a pro-worker and pro-social site of data work be transferred to other sites of data production? Said differently, if we know an alternative is possible, how can that improved social dynamic be developed at other sites? Adhering to the cosmolocal approach, each site of potential transfer will have to be studied as an ecosystem of data work production, with attention to the power dynamics and relationships that exist between workers and requesters. Subsequent intervention to (re)formulate and instill a more trusting, mutually beneficial relationship between workers and requesters must be tailored to those mappings. For example, beginning the process of transfer to Amazon MTurk should take into account the *triangular* relationship between stakeholders (Amazon–as platform, workers–Turkers, and requesters) [416]). For other, less well known platforms these dynamics are less clear. Consequently, a starting point might be mapping the communication flows on these platforms, which, in conjunction with

interviews with both workers and requesters, could provide context for the formulation of power dynamics on the respective platforms.

Creating quality datasets is also relevant to the increasingly prominent paradigm of *frugal AI* [433, 434]. The emphasis with frugal AI is to create systems that meet or exceed the current status quo, but use far less resources to do so. The recent model release (R1) by DeepSeek [435] appears to follow this trend. If we consider using less, but better data, the role of data worker-as-auditor continues to be an important one. Further, while there are not yet well-established specifications for data used to train and refine systems developed with the frugal AI approach, though some early work points to methods generally used in environments with limited data availability [436], which again, points to the importance of data quality vs data quantity.

6.4.2 Part 2: Responding to the evolving nature of data work

Concurrent with the new national emphasis on AI (at time of writing) [437, 438, 439] is a push to make the process of AI development even less "human". These changes affect both the data and data workers necessary to the creation of AI systems.

There has been a push among AI researchers to use LLMs to produce data annotation in place of human data workers [440, 441, 442]. There has been work demonstrating that this (replacing humans with LLMs) is neither feasible nor desirable. Specifically, the generation of such synthetic data causes issues regarding representation [443, 444] and model collapse more broadly [445, 446, 447]. As it appears neither side has intellectual dominance within the larger research community yet, it is likely that data workers will be increasingly working in tandem with AI tools.

Instead of generating and labeling initial data, data workers are much more likely to be in the position of refining systems that will be in use with (or to determine things about) human populations, e.g., [448]. Further, they will likely need to spot when AI systems

are going "MAD"[1] ([449]) which requires a level of systems-savvy that is beyond general knowledge. Subsequently, the critical data literacy curriculum I introduce in this dissertation should be updated to adapt to this increasingly supervisory (vs creator) role.

[1] "MAD" stands for Model Autophagy Disorder, a phenomenon the authors describe as akin to Mad Cow Disease, in which next-generation models are trained off of synthetic data in an autophagous cycle.

CHAPTER 7
CONCLUSION

This dissertation explores what the role of data workers' perspective and lived experience, when applied to their work with datasets, can do to change the paradigm of AI (and other data-intensive) system development. I premise this work on the idea that data workers are well-positioned to act as dataset auditors; it is data workers who review, annotate, and collate the individual constituent datum of the massive datasets used to train and refine AI systems. Data workers can thus alert us to concerning entries in a dataset, where concerning could be in terms of content (for example, harmful or offensive) or unexpected inclusion, given the dataset's specifications (e.g., a photo exclusively of a dog in a dataset of cats).

But why do we need dataset auditors? Returning to my research questions (1.2), each chapter of this dissertation represents a subset of work that relates to a facet of *what is the role of perspective in data work, and how can we incorporate the perspective of data workers as partners in dataset contextualization?*. The first half of this dissertation is devoted to my first research question, or understanding what measures are necessary, in terms of workplace culture and worker training, to create the space and environment in which data workers are able to generate and share their dataset understandings.

The first content chapter explains my personal motivation for engaging in this work, namely the first-hand observation of how a lack of personal agency and relationship to a data-intensive, generalizable system manifests in the quality of data work produced. Specifically, I report on an early client project at DataWorks, where the data workers were not producing work that met the client's expectation, not for lack of trying. There were two problems at play: first, workers didn't understand how the data annotations they produced were going to be used (and, subsequently, what format they needed to take), and second, was the system was being trained for, and therefore, what their annotations needed to focus

on. These mistakes, or misunderstandings, are indicative of a larger trend in AI and ML, in which data annotators are held apart from system ideation and development, and thus disconnected from the larger goals of the project.

When we see similar problems taking place with other datasets, particularly massive, critical benchmarking ones, such as *ImageNet* and *WordNet*, I theorize that similar problems are very likely at play. When data annotators lack both the agency to raise alarm, and the knowledge of how to translate their concerns to the data work requesters, their perspective goes unheard. I then explain how this dynamic exists (if even more exacerbated) also on a larger scale in quasi-anonymous platform data work.

In the second content chapter, I discuss the impact on data work when these two things (agency and insider knowledge) can be confidently practiced by data workers. To help develop both, I designed a workplace training in CDL for the Data Fellows at DataWorks, to help them leverage their existing perspectives and experiences and to help them channel those things into their data work. The workshop was not so much about how, e.g., ML works, but about the role that the data workers play in that system or pipeline and how the requester(s) of their labor understood the work they produced. We saw the results of this training coming into play in the following section, where I shared I case study of a project for a NLP client of DataWorks. The Data Fellows were able to articulate their concerns to the project client, based on their own knowledge, experiences, and perspective, but with the language and understanding of ML systems to back them up and assist with explaining why the scale of this project (the NLP system) made these concerns all the more pressing.

While this project did not take place in a spreadsheet, it was kind the of project that often does—as a tagging and translation project—and instead of the data workers having to flag entries by writing down task numbers and passing those along over email, we can imagine that it would be far more convenient to flag the entries in situ, on spreadsheet itself. This is a large part of the rationale behind the tool that I describe in the final section of chapter, *Datum Fieldnotes*. This Google Sheets add-on supports the integration and

archiving of perspectives into spreadsheet-based data work, namely that taking place in Google Sheets. As a tool, the Fieldnotes supports a critical lens on data, by allowing data workers to annotate individual, cell-level changes with comments, questions, and reflection. Further, the log makes it both possible to investigate historical changes to the spreadsheet, giving the 'who, what, where, when, and how' that support critical data literacy [12].

In the third content chapter, I step back to look more broadly at how data work happens on a large scale, given that DataWorks is a very unusual site by typical data work standards. In focusing on quasi-anonymous data work platforms, I am studying sites where there are triangular relationships between platforms, requesters, and workers, but information flow is not bi-directional; in particular, workers are told to produce data (as task submissions) but given little avenue for reflection or concern raising about what they see in those tasks. Further, there is rarely much discussion about what the point of labeling or annotating a given dataset is, besides the fact that the requester wants it done and is willing to pay. Therefore, many data workers on these platforms don't know what kinds of projects they are contributing to, not are they seen as collaborators on these projects, despite providing essential data that makes them possible.

In this chapter, I begin with how we can create more pro-social working conditions for platform data workers, as academic and industry requesters. The second piece of work in this chapter extends this, examining how requesters understand workers at large, particularly requesters of data labor that is used to train or refine AI and ML systems. In that work, I find that requesters, given that they usually know nothing about the workers, tend to treat workers hostilely, looking for proof that they hold certain characteristics and that they will provide a meticulous quality of work. In order to ascertain these qualities, they employ a number of quantitative *proxies*, or automated tests of numeric values they do have access to, with the hopes of employing those to make sense of qualitative concerns, such as whether or not the worker produced a high quality submission. These proxies miss two things. First, the proxies are often not founded in evidence-based methods, instead

they are arbitrary and anecdotal. Second, they neglect that skill and expertise can upend these quantitative measures, instead producing undesirable values, such as a highly skilled worker moving through a task faster than their peers.

The final content chapter of this dissertation is focused on how the learnings of DataWorks, specifically related to critical data literacy and workplace autonomy & engaged workplace (as necessary components to enlist data workers as dataset auditors) can be transferred to quasi-anonymous platforms. Here, I invoke the concept of the *cosmolocal* [159] to reiterate that the DataWorks model is a unique one, and that it should not be used as an inspiration for a one-to-one transfer. Instead, we should treat DataWorks as a source of knowledge that can be applied to other sites, in terms of developing the pre-conditions for data worker-as-auditor, while respecting the nuance and context of the transferee locale.

Overall, this proposal is in the advancement of bringing more—in both quantity and variety—perspectives into data work, with an emphasis on employing these perspectives (implausible without forming respectful, collaborative relationships with data workers) as a means of re-localizing and re-contextualizing datasets that have been denuded of anything not generalizable.

Appendices

APPENDIX A
ACRONYMS AND RELEVANT TERMS

In this section, I list terms that are used throughout this proposal.

Critical data literacy (CDL) — termed by Tygel & Kirsch, CDL is an expansion on data literacy in which those literacy skills are used for emancipation, including contextual interpretation, questioning of common sense concepts, and development of new concepts [12].

Crowdworking — the distribution of large tasks to a large number of workers, usually via a platform like Amazon Mechanical Turk (Amazon MTurk) [81, 88]. For this proposal, crowdworking is assumed to take place via a digital platform, and the platforms in question are those that are either primarily for data work tasks, or data work forms a large percentage of the available tasks.

Data literacy — the ability to read and make sense of information represented numerically (e.g., percentages and averages), which can be employed algorithmically for inference, and as well as being able to read and make sense of visual representations of such information [450].

Data work — the collection and preparation of data to be used for analysis or algorithmic systems, including standardization, organization, labeling, and annotation [73, 365].

(Quasi-anonymous) Digital data work platforms — web-based platforms on which requesters of data labor can post tasks to be completed by workers, who depending on the given platform, are often globally distributed [451, 22, 51]. When the worker and requester, by default, know almost nothing about one another, the platform has an "anonymous" model, however work has shown many such models are actually easily to abuse and de-anonymize workers [104], thus they are "quasi-anonymous".

REFERENCES

[1] S. Sirimanne, "How artificial intelligence chatbots could affect jobs | UNCTAD," *United Nations Conference on Trade and Development*, Jan. 2023.

[2] M. Kan, "OpenAI Surveying People on 'Economic Impact' of ChatGPT," *PCMAG*, Feb. 2023.

[3] A. Agrawal, J. Gans, and A. Goldfarb, "ChatGPT and How AI Disrupts Industries," *Harvard Business Review*, Dec. 2022, Section: AI and machine learning.

[4] S. Gilbert, *What might be the economic impact of AI tools like ChatGPT?* Jun. 2023.

[5] D. Rothman, "ChatGPT is about to revolutionize the economy. We need to decide what that looks like.," *MIT Technology Review*, Mar. 2023.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, ISSN: 1063-6919, Jun. 2009, pp. 248–255.

[7] G. A. Miller, "WordNet: A lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[8] A. Birhane and V. U. Prabhu, "Large image datasets: A pyrrhic win for computer vision?" In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, ISSN: 2642-9381, Jan. 2021, pp. 1536–1546.

[9] C. DiSalvo, A. Rothschild, L. L. Schenck, B. R. Shapiro, and B. DiSalvo, "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*, vol. 8, no. CSCW1, p. 23, Jan. 2023.

[10] T. Koltay, "Data literacy: In search of a name and identity," *Journal of Documentation*, vol. 71, no. 2, pp. 401–415, Mar. 2015.

[11] P. Freire, *Pedagogy of the oppressed*, 30th anniversary ed. New York: Continuum, 2000, ISBN: 978-0-8264-1276-8.

[12] A. F. Tygel and R. Kirsch, "Contributions of Paulo Freire for a critical data literacy: A popular education approach," *The Journal of Community Informatics*, vol. 12, no. 3, 2016.

[13] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, ""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*, ser. CHI '21, New York, NY, USA: Association for Computing Machinery, May 2021, pp. 1–15, ISBN: 978-1-4503-8096-6.

[14] R. Kitchin, *Critical data studies: an A to Z guide to concepts and methods*, First edition. Hoboken, NJ: Polity press, 2025, ISBN: 978-1-5095-6652-5 978-1-5095-6653-2.

[15] M. Miceli, M. Schuessler, and T. Yang, "Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, pp. 1–25, Oct. 2020.

[16] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ISSN: 2640-3498, PMLR, Jan. 2018, pp. 77–91.

[17] K. Hill, "Wrongfully Accused by an Algorithm," *The New York Times*, Jun. 2020.

[18] D. Tsipras, S. Santurkar, L. Engstrom, A. Ilyas, and A. Madry, "From ImageNet to Image Classification: Contextualizing Progress on Benchmarks," in *Proceedings of the 37th International Conference on Machine Learning*, ISSN: 2640-3498, PMLR, Nov. 2020, pp. 9625–9635.

[19] A. Birhane, V. U. Prabhu, and E. Kahembwe, "Multimodal datasets: Misogyny, pornography, and malignant stereotypes," *arXiv:2110.01963 [cs]*, Oct. 2021, arXiv: 2110.01963.

[20] M. Miceli, J. Posada, and T. Yang, "Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power?" *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. GROUP, pp. 1–14, Jan. 2022.

[21] M. Miceli *et al.*, "Documenting data production processes: A participatory approach for data work," *Proc. ACM Hum.-Comput. Interact.*, vol. 6, no. CSCW2, Nov. 2022, Number of pages: 34 Place: New York, NY, USA Publisher: Association for Computing Machinery tex.articleno: 510 tex.issue_date: November 2022.

[22] M. L. Gray and S. Suri, *Ghost work: how to stop Silicon Valley from building a new global underclass*. Boston: Houghton Mifflin Harcourt, 2019, ISBN: 978-1-328-56628-7.

[23] A. Alkhatib, M. S. Bernstein, and M. Levi, "Examining Crowd Work and Gig Work Through The Historical Lens of Piecework," in *Proceedings of the 2017 CHI Con-*

ference on Human Factors in Computing Systems, Denver Colorado USA: ACM, May 2017, pp. 4599–4616, ISBN: 978-1-4503-4655-9.

[24] J. Browne, E. Drage, and K. McInerney, "Tech workers' perspectives on ethical issues in AI development: Foregrounding feminist approaches," *Big Data & Society*, vol. 11, no. 1, p. 20 539 517 231 221 780, Mar. 2024, Publisher: SAGE Publications Ltd.

[25] C. K. Lin and S. J. Jackson, "From Bias to Repair: Error as a Site of Collaboration and Negotiation in Applied Data Science Work," *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, no. CSCW1, 131:1–131:32, Apr. 2023.

[26] R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, and P. Hall, "Towards a standard for identifying and managing bias in artificial intelligence," National Institute of Standards and Technology (U.S.), Gaithersburg, MD, Tech. Rep. NIST SP 1270, Mar. 2022, NIST SP 1270.

[27] R. Benjamin, *Race after technology: abolitionist tools for the new Jim code*. Medford, MA: Polity, 2019, ISBN: 978-1-5095-2643-7.

[28] E. Sheppard, "The Spaces and Times of Globalization: Place, Scale, Networks, and Positionality*," *Economic Geography*, vol. 78, no. 3, pp. 307–330, 2002, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1944-8287.2002.tb00189.x.

[29] S. B. Merriam, J. Johnson-Bailey, M.-Y. Lee, Y. Kee, G. Ntseane, and M. Muhamad, "Power and positionality: Negotiating insider/outsider status within and across cultures," *International Journal of Lifelong Education*, vol. 20, no. 5, pp. 405–416, Sep. 2001, Publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1080/02601370

[30] M. Lengwiler, "Participatory Approaches in Science and Technology: Historical Origins and Current Practices in Critical Perspective," *Science, Technology, & Human Values*, vol. 33, no. 2, pp. 186–200, Mar. 2008, Publisher: SAGE Publications Inc.

[31] D. Haraway, "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective," *Feminist Studies*, vol. 14, no. 3, p. 575, 1988.

[32] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias," *ProPublica*, May 2016.

[33] R. L. Brooks, "Critical Race Theory: A Proposed Structure and Application to Federal Pleading," *Harvard Blackletter Law Journal*, vol. 11, p. 85, 1994.

[34] G. Ladson-Billings and W. F. Tate IV, "Toward a Critical Race Theory of Education," *Teachers College Record*, vol. 97, no. 1, p. 23, 1995.

[35] S. Samuel, "Why it's so damn hard to make AI fair and unbiased," *Vox*, Apr. 2022.

[36] N. Parvin and A. Pollock, "Unintended by Design: On the Political Uses of "Unintended Consequences"," *Engaging Science, Technology, and Society*, vol. 6, p. 320, Aug. 2020.

[37] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21, New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 610–623, ISBN: 978-1-4503-8309-7.

[38] L. Suchman, "The uncontroversial 'thingness' of AI," *Big Data & Society*, vol. 10, no. 2, p. 20 539 517 231 206 794, Jul. 2023, Publisher: SAGE Publications Ltd.

[39] S. Paasonen, "From Cybernation to Feminization: Firestone and Cyberfeminism," in *Further Adventures of the Dialectic of Sex*, M. Merck and S. Sandford, Eds., New York: Palgrave Macmillan US, 2010, pp. 61–83, ISBN: 978-1-349-28650-8 978-0-230-10999-5.

[40] R. Braidotti, "Cyberfeminism with a difference," *New Formations*, vol. 29, Jan. 1996.

[41] P. Pressau, "Cyberfeminism : Encyclopedia of New Media,"

[42] A. Everett, "On Cyberfeminism and Cyberwomanism: High-Tech Mediations of Feminism's Discontents," *Signs: Journal of Women in Culture and Society*, vol. 30, no. 1, pp. 1278–1286, Sep. 2004, Publisher: The University of Chicago Press.

[43] S. R. Schulte, "Surfing Feminism's Online Wave: The Internet and the Future of Feminism," *Feminist Studies*, vol. 37, no. 3, pp. 727–744, 2011, Publisher: Feminist Studies, Inc.

[44] S. Paasonen, "Revisiting cyberfeminism," vol. 36, no. 3, pp. 335–352, Sep. 2011, Publisher: De Gruyter Mouton Section: Communications.

[45] M. Tazi and K. Oumlil, "The Rise of Fourth-Wave Feminism in the Arab region? Cyberfeminism and Women's Activism at the Crossroads of the Arab Spring," *CyberOrient*, vol. 14, no. 1, pp. 44–71, 2020, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/

[46] S. Gillis, G. Howie, and R. Munford, Eds., *Third Wave Feminism*. London: Palgrave Macmillan UK, 2007, ISBN: 978-0-230-52174-2 978-0-230-59366-4.

[47] C. D'Ignazio and L. F. Klein, *Data feminism* (Strong ideas series). Cambridge, Massachusetts: The MIT Press, 2020, ISBN: 978-0-262-04400-4.

[48] L. Klein and C. D'Ignazio, "Data Feminism for AI," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '24, New York, NY, USA: Association for Computing Machinery, Jun. 2024, pp. 100–112, ISBN: 979-8-4007-0450-5.

[49] M. G. Crain, W. R. Poster, and M. A. Cherry, Eds., *Invisible Labor: Hidden Work in the Contemporary World*, 1st ed. University of California Press, 2016, ISBN: 978-0-520-28640-5.

[50] S. Darian *et al.*, "Enacting Data Feminism in Advocacy Data Work," *Proc. ACM Hum.-Comput. Interact.*, vol. 7, no. CSCW1, 47:1–47:28, Apr. 2023.

[51] M. Miceli and J. Posada, "The Data-Production Dispositif," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, 460:1–460:37, Nov. 2022.

[52] "Coffee Roasters' Data Vernacular: On the Entanglement of Digital Data and Craft," in *Situating Data: Inquiries in Algorithmic Culture*, Nieuwe Prinsengracht 89 1018 VR Amsterdam Nederland: Amsterdam University Press, Feb. 2023, ISBN: 978-90-485-5544-4 978-94-6372-297-1.

[53] A. L. Tsing, "On Nonscalability," *Common Knowledge*, vol. 18, no. 3, pp. 505–524, Aug. 2012.

[54] "The Cosmolocal Reader," P2P Foundation, Tech. Rep.

[55] P. A. Loring, "A vernacular for living systems: Alternative framings for the future of food," *Futures*, vol. 154, p. 103 276, Dec. 2023.

[56] S. Amrute, R. Singh, and R. L. Guzmán, *A Primer on AI in/from the Majority World: An Empirical Site and a Standpoint*, SSRN Scholarly Paper, Rochester, NY, Sep. 2022.

[57] I. Huvila, *Perspectives on Paradata: Research and Practice of Documenting Process Knowledge* (Knowledge Management and Organizational Learning Series v.13), 1st ed. Cham: Springer International Publishing AG, 2024, ISBN: 978-3-031-53946-6.

[58] C. B. Trace and J. A. Hodges, "The Role of Paradata in Algorithmic Accountability," in *Perspectives on Paradata: Research and Practice of Documenting Process Knowledge*, I. Huvila, L. Andersson, and O. Sköld, Eds., Cham: Springer International Publishing, 2024, pp. 197–213, ISBN: 978-3-031-53946-6.

[59] D. U. Patton, D.-W. Brunton, A. Dixon, R. J. Miller, P. Leonard, and R. Hackman, "Stop and Frisk Online: Theorizing Everyday Racism in Digital Policing in the Use

of Social Media for Identification of Criminal Conduct and Associations," *Social Media + Society*, vol. 3, no. 3, p. 205 630 511 773 334, Jul. 2017.

[60] A. E. Marwick and d. boyd, "I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience," *New Media & Society*, vol. 13, no. 1, pp. 114–133, Feb. 2011, Publisher: SAGE Publications.

[61] C. Fieseler, E. Bucher, and C. P. Hoffmann, "Unfairness by Design? The Perceived Fairness of Digital Labor on Crowdworking Platforms," *Journal of Business Ethics*, vol. 156, no. 4, pp. 987–1005, Jun. 2019.

[62] K. Börner, "Data Visualization Literacy," in *Proceedings of the 27th ACM Conference on Hypertext and Social Media*, Halifax Nova Scotia Canada: ACM, Jul. 2016, pp. 1–1, ISBN: 978-1-4503-4247-6.

[63] B. Johnson, B. Rydal Shapiro, B. DiSalvo, A. Rothschild, and C. DiSalvo, "Exploring approaches to data literacy through a critical race theory perspective," in *Proceedings of the 2021 CHI conference on human factors in computing systems*, 2021, pp. 1–15.

[64] C. G. Northcutt, A. Athalye, and J. Mueller, *Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks*, arXiv:2103.14749 [cs, stat], Nov. 2021.

[65] *Label Errors in Benchmark ML Datasets*.

[66] F. F. Li, *The Worlds I See: Curiosity, Exploration, and Discovery at the Dawn of AI*. Flatiron Books: A Moment of Lift Book, 2023.

[67] S. Levy, "Fei-Fei Li Started an AI Revolution by Seeing Like an Algorithm," *Wired*, Section: tags.

[68] A. Barr, *Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms*, Newspaper, Jul. 2015.

[69] *UCD student's research takes down an 80-million image artificial intelligence database*, Section: News, Jul. 2020.

[70] *80 Million Tiny Images*.

[71] S. Kapania, D. Wang, and A. S. Taylor, "Annotator Diversity in Data Practices," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2023.

[72] J. Nicas, "Atlanta Asks Google Whether It Targeted Black Homeless People," *The New York Times*, Oct. 2019.

[73] N. Sambasivan and R. Veeraraghavan, "The Deskilling of Domain Expertise in AI Development," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI '22, New York, NY, USA: Association for Computing Machinery, Apr. 2022, pp. 1–14, ISBN: 978-1-4503-9157-3.

[74] M. Muller *et al.*, "Human-Centered Study of Data Science Work Practices," in *Extended Abstracts of the 2019 CHI Conference*, 2019, p. 8.

[75] M. Kogan, A. Halfaker, S. Guha, C. Aragon, M. Muller, and S. Geiger, "Mapping Out Human-Centered Data Science: Methods, Approaches, and Best Practices," in *Companion of the 2020 ACM International Conference on Supporting Group Work*, Sanibel Island Florida USA: ACM, Jan. 2020, pp. 151–156, ISBN: 978-1-4503-6767-7.

[76] S. Chancellor, "Toward Practices for Human-Centered Machine Learning," *Communications of the ACM*, vol. 66, no. 3, pp. 78–85, Feb. 2023.

[77] M. Muller *et al.*, "How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow Scotland Uk: ACM, May 2019, pp. 1–15, ISBN: 978-1-4503-5970-2.

[78] L. Nader, "Up the Anthropologist: Perspectives Gained From Studying Up," Tech. Rep., 1972, ERIC Number: ED065375.

[79] L. C. Irani and M. S. Silberman, "Turkopticon: Interrupting worker invisibility in amazon mechanical turk," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Paris France: ACM, Apr. 2013, pp. 611–620, ISBN: 978-1-4503-1899-0.

[80] L. C. Irani and M. S. Silberman, "Stories We Tell About Labor: Turkopticon and the Trouble with "Design"," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, San Jose California USA: ACM, May 2016, pp. 4573–4586, ISBN: 978-1-4503-3362-7.

[81] G. Jäger, L. S. Zilian, C. Hofer, and M. Füllsack, "Crowdworking: Working with or against the crowd?" *Journal of Economic Interaction and Coordination*, vol. 14, no. 4, pp. 761–788, Dec. 2019.

[82] M. E. Whiting *et al.*, "Crowd Guilds: Worker-led Reputation and Feedback on Crowdsourcing Platforms," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, Portland Oregon USA: ACM, Feb. 2017, pp. 1902–1913, ISBN: 978-1-4503-4335-0.

[83] W. Hawkins and B. Mittelstadt, "The ethical ambiguity of AI data enrichment: Measuring gaps in research ethics norms and practices," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '23, New York, NY, USA: Association for Computing Machinery, Jun. 2023, pp. 261–270, ISBN: 9798400701924.

[84] M. K. Scheuerman, M. Pape, and A. Hanna, "Auto-essentialization: Gender in automated facial analysis as extended colonial project," *Big Data & Society*, vol. 8, no. 2, p. 205 395 172 110 537, Jul. 2021.

[85] Y. A. Loukissas, *All data are local: thinking critically in a data-driven society*. Cambridge, Massachusetts: The MIT Press, 2019, ISBN: 978-0-262-03966-6.

[86] L. Stark and J. Hutson, "Physiognomic Artificial Intelligence," *Fordham Intellectual Property, Media and Entertainment Law Journal*, vol. 32, no. 4, p. 922, Jan. 2022.

[87] T. DeVries, I. Misra, and C. Wang, "Does Object Recognition Work for Everyone?,"

[88] P. Jones, *Work without the worker: labour in the age of platform capitalism*. Brooklyn: Verso Books, 2021, ISBN: 978-1-83976-043-3.

[89] S. T. Roberts, *Behind the screen: content moderation in the shadows of social media*. New Haven: Yale University Press, 2019, OCLC: on1055263168, ISBN: 978-0-300-23588-3.

[90] C. Du and M. Lease, *Voices of Workers: Why a Worker-Centered Approach to Crowd Work Is Challenging*, arXiv:2212.14471 [cs], Jan. 2023.

[91] D. Wang, S. Prabhat, and N. Sambasivan, "Whose AI Dream? In search of the aspiration in data annotation.," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI '22, New York, NY, USA: Association for Computing Machinery, Apr. 2022, pp. 1–16, ISBN: 978-1-4503-9157-3.

[92] S. Kapania, A. S. Taylor, and D. Wang, "A hunt for the Snark: Annotator Diversity in Data Practices," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg Germany: ACM, Apr. 2023, pp. 1–15, ISBN: 978-1-4503-9421-5.

[93] J. Oppenlaender, K. Milland, A. Visuri, P. Ipeirotis, and S. Hosio, "Creativity on Paid Crowdsourcing Platforms," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA: ACM, Apr. 2020, pp. 1–14, ISBN: 978-1-4503-6708-0.

[94] M. E. Whiting, G. Hugh, and M. S. Bernstein, "Fair Work: Crowd Work Minimum Wage with One Line of Code," *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, no. 1, pp. 197–206, Oct. 2019, Section: Technical Papers.

[95] R. C. Gomer and E. Simperl, "Trusts, co-ops, and crowd workers: Could we include crowd data workers as stakeholders in data trust design?" *Data & Policy*, vol. 2, e20, Jan. 2020, Publisher: Cambridge University Press.

[96] D. Birnbaum and M. Somers, "Past as prologue: Taylorism, the new scientific management and managing human capital," *International Journal of Organizational Analysis*, vol. 31, no. 6, pp. 2610–2622, 2023, Publisher: Emerald Publishing Limited.

[97] A. M. Brawley and C. L. Pury, "Work experiences on MTurk: Job satisfaction, turnover, and information sharing," *Computers in Human Behavior*, vol. 54, pp. 531–546, Jan. 2016.

[98] B. V. Hanrahan, A. Chen, J. Ma, N. F. Ma, A. Squicciarini, and S. Savage, "The Expertise Involved in Deciding which HITs are Worth Doing on Amazon Mechanical Turk," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, 128:1–128:23, Apr. 2021.

[99] C.-W. Chiang, A. Kasunic, and S. Savage, "Crowd Coach: Peer Coaching for Crowd Workers' Skill Growth," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, 37:1–37:17, Nov. 2018.

[100] P. Popiel, ""Boundaryless" in the creative economy: Assessing freelancing on Upwork," *Critical Studies in Media Communication*, vol. 34, no. 3, pp. 220–233, May 2017, Publisher: Routledge _eprint: https://doi.org/10.1080/15295036.2017.1282618.

[101] J. Alvarez De La Vega, M. Cecchinato, and J. Rooksby, ""Why lose control?" A Study of Freelancers' Experiences with Gig Economy Platforms," in Y. Kitamura, A. Quigley, K. Isbister, T. Igarashi, P. Bjørn, and S. Drucker, Eds., Conference Name: ACM CHI Conference on Human Factors in Computing Systems 2021 Meeting Name: ACM CHI Conference on Human Factors in Computing Systems 2021, New York: ACM, May 2021, p. 455, ISBN: 978-1-4503-8096-6.

[102] E. Foong, "Understanding and Designing Sociotechnical Systems to Support the Impression Management Practices of Online Freelance Workers," in *Companion of the 2020 ACM International Conference on Supporting Group Work*, Sanibel Island Florida USA: ACM, Jan. 2020, pp. 25–33, ISBN: 978-1-4503-6767-7.

[103] D. Howcroft and B. Bergvall-Kåreborn, "A Typology of Crowdwork Platforms," *Work, Employment and Society*, vol. 33, no. 1, pp. 21–38, Feb. 2019, Publisher: SAGE Publications Ltd.

[104] M. Lease *et al.*, *Mechanical Turk is Not Anonymous*, SSRN Scholarly Paper, Rochester, NY, Mar. 2013.

[105] D. Gershgorn, *The data that transformed AI research—and possibly the world*, Jul. 2017.

[106] S. Sannon, "Power Dynamics and Privacy Protection in the Gig Economy," in *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, Austin TX USA: ACM, Nov. 2019, pp. 84–88, ISBN: 978-1-4503-6692-2.

[107] S. Sannon and D. Cosley, "Privacy, Power, and Invisible Labor on Amazon Mechanical Turk," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow Scotland Uk: ACM, May 2019, pp. 1–12, ISBN: 978-1-4503-5970-2.

[108] S. Sannon and D. Cosley, ""It was a shady HIT": Navigating Work-Related Privacy Concerns on MTurk," in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal QC Canada: ACM, Apr. 2018, pp. 1–6, ISBN: 978-1-4503-5621-3.

[109] S. Uzor, J. T. Jacques, J. J. Dudley, and P. O. Kristensson, "Investigating the Accessibility of Crowdwork Tasks on Mechanical Turk," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, May 2021, pp. 1–14, ISBN: 978-1-4503-8096-6.

[110] A. Vashistha, P. Sethi, and R. Anderson, "BSpeak: An Accessible Voice-based Crowdsourcing Marketplace for Low-Income Blind People," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal QC Canada: ACM, Apr. 2018, pp. 1–13, ISBN: 978-1-4503-5620-6.

[111] P. Narula, P. Gutheim, D. Rolnitzky, A. Kulkarni, and B. Hartmann, "MobileWorks: A Mobile Crowdsourcing Platform for Workers at the Bottom of the Pyramid," in *Proceedings of the 11th AAAI Conference on Human Computation*, AAAI Press, 2011, pp. 121–123.

[112] L. Lascau, S. J. J. Gould, A. L. Cox, E. Karmannaya, and D. P. Brumby, "Monotasking or Multitasking: Designing for Crowdworkers' Preferences," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow Scotland Uk: ACM, May 2019, pp. 1–14, ISBN: 978-1-4503-5970-2.

[113] M. Yin, S. Suri, and M. L. Gray, "Running Out of Time: The Impact and Value of Flexibility in On-Demand Crowdwork," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal QC Canada: ACM, Apr. 2018, pp. 1–11, ISBN: 978-1-4503-5620-6.

[114] P. Beytía and C. Wagner, "Visibility layers: A framework for systematising the gender gap in Wikipedia content," *Internet Policy Review*, vol. 11, no. 1, Mar. 2022.

[115] A. James, "Women in the gig economy: Feminising 'digital labour'," *Work in the Global Economy*, pp. 1–25, Mar. 2022.

[116] M. Hicks, *Programmed inequality: how Britain discarded women technologists and lost its edge in computing* (History of computing). Cambridge, MA: MIT Press, 2017, ISBN: 978-0-262-03554-5.

[117] M. W. Davies, *Woman's Place Is at the Typewriter: Office Work and Office Workers, 1870-1930*. Temple University Press, Nov. 2018, ISBN: 978-1-4399-1817-3.

[118] T. J. Misa, "Masculinity and the Machine Man," in *Gender Codes: Why Women Are Leaving Computing*, Conference Name: Gender Codes: Why Women Are Leaving Computing, IEEE, 2010, pp. 51–71, ISBN: 978-0-470-61991-9.

[119] R. E. Weber and B. Gilchrist, "Discrimination in the employment of women in the computer industry," *Communications of the ACM*, vol. 18, no. 7, pp. 416–418, Jul. 1975.

[120] F. Jahanbakhsh, J. Cranshaw, S. Counts, W. S. Lasecki, and K. Inkpen, "An Experimental Study of Bias in Platform Worker Ratings: The Role of Performance Quality and Gender," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA: ACM, Apr. 2020, pp. 1–13, ISBN: 978-1-4503-6708-0.

[121] M. Graham *et al.*, "The Fairwork Foundation: Strategies for Improving Platform Work," *Weizenbaum Conference*, 2019, Publisher: WI - Weizenbaum Institute for the Networked Society Version Number: 1.

[122] C. Gerber, "Community building on crowdwork platforms: Autonomy and control of online workers?" *Competition & Change*, vol. 25, no. 2, pp. 190–211, Apr. 2021, Publisher: SAGE Publications Ltd.

[123] X. Wang, H. Zhu, Y. Li, Y. Cui, and J. Konstan, "A Community Rather Than A Union: Understanding Self-Organization Phenomenon on MTurk and How It Impacts Turkers and Requesters," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, Denver Colorado USA: ACM, May 2017, pp. 2210–2216, ISBN: 978-1-4503-4656-6.

[124] J. Bates, A. Checco, and E. Gerakopoulou, "Worker Perspectives on Designs for a Crowdwork Co-operative," in *New Perspectives in Critical Data Studies: The Ambivalences of Data Power*, ser. Transforming Communications – Studies in Cross-Media Research, Cham: Springer International Publishing, 2022, ISBN: 978-3-030-96179-4 978-3-030-96180-0.

[125] *End the Harm of Mass Rejections*, May 2023.

[126] *Mass rejections - Turkopticon*, Section: Articles, Dec. 2020.

[127] W. van Zoonen, A. E. Sivunen, and J. W. Treem, "Algorithmic management of crowdworkers: Implications for workers' identity, belonging, and meaningfulness of work," *Computers in Human Behavior*, vol. 152, p. 108 089, Mar. 2024.

[128] M. Steiger, T. J. Bharucha, S. Venkatagiri, M. J. Riedl, and M. Lease, "The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, May 2021, pp. 1–14, ISBN: 978-1-4503-8096-6.

[129] *Our Foundation Story*.

[130] B. Perrigo, "Meta Accused Of Human Trafficking and Union-Busting in Kenya," *TIME*, May 2022.

[131] B. Perrigo, "Inside Facebook's African Sweatshop," *TIME*, Feb. 2022.

[132] K. Hao and D. Seetharaman, "Cleaning Up ChatGPT Takes Heavy Toll on Human Workers," *Wall Street Journal*, Jul. 2023.

[133] N. Rowe, "'It's destroyed me completely': Kenyan moderators decry toll of training of AI models," *The Guardian*, Aug. 2023.

[134] N. Rowe, "Underage Workers Are Training AI," *Wired*, Section: tags.

[135] P. Hitlin, "Research in the Crowdsourcing Age, a Case Study," Pew Research Center, Tech. Rep., Jul. 2016, p. 37.

[136] V. A. Rivera and D. T. Lee, "I Want to, but First I Need to: Understanding Crowdworkers' Career Goals, Challenges, and Tensions," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, 150:1–150:22, Apr. 2021.

[137] A. Idowu and A. Elbanna, "Crowdworkers, social affirmation and work identity: Rethinking dominant assumptions of crowdwork1," *Information and Organization*,

New Ways of Working: Rematerializing Organization in the Digital Age, vol. 31, no. 4, p. 100 335, Dec. 2021.

[138] Y. Sun, X. Ma, K. Ye, and L. He, "Investigating Crowdworkers' Identify, Perception and Practices in Micro-Task Crowdsourcing," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. GROUP, 35:1–35:20, Jan. 2022.

[139] S. Savage, C. W. Chiang, S. Saito, C. Toxtli, and J. Bigham, "Becoming the Super Turker:Increasing Wages via a Strategy from High Earning Workers," in *Proceedings of The Web Conference 2020*, Taipei Taiwan: ACM, Apr. 2020, pp. 1241–1252, ISBN: 978-1-4503-7023-3.

[140] H. Xia, "Undue Influence or Exploitation — A Qualitative Inquiry into an Ethical Dilemma Around Payment in Crowd Work-Based Research in the U.S.," *Computer Supported Cooperative Work (CSCW)*, May 2023.

[141] L. K. Howe, *Pink collar workers: inside the world of women's work*. New York: Putnam, 1977.

[142] D. A. Grier, *When computers were human*, 1. paperback print. Princeton, N.J.: Princeton Univ. Press, 2007, ISBN: 978-0-691-13382-9.

[143] D. G. Widder and R. Wong, *Thinking Upstream: Ethics and Policy Opportunities in AI Supply Chains*, arXiv:2303.07529 [cs], Mar. 2023.

[144] D. G. Widder and D. Nafus, "Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility," *Big Data & Society*, vol. 10, no. 1, p. 20 539 517 231 177 620, Jan. 2023, Publisher: SAGE Publications Ltd.

[145] J. Zong and J. N. Matias, "Data Refusal From Below: A Framework for Understanding, Evaluating, and Envisioning Refusal as Design," *ACM Journal on Responsible Computing*, Oct. 2023, Just Accepted.

[146] B. Honig, *A Feminist Theory of Refusal*. Harvard University Press, 2021, ISBN: 978-0-674-24849-6.

[147] *Feminist Data Manifest-No*.

[148] R. Y. Wong, "Tactics of Soft Resistance in User Experience Professionals' Values Work," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW2, 355:1–355:28, Oct. 2021.

[149] J. Muldoon, M. Graham, and C. Cant, *Feeding the machine: the hidden human labour powering AI*. Edinburgh: Canongate, 2024, ISBN: 978-1-83726-181-9 978-1-83726-182-6.

[150] J. Muldoon, C. Cant, M. Graham, and F. Ustek Spilda, "The poverty of ethical AI: Impact sourcing and AI supply chains," *AI & SOCIETY*, Dec. 2023.

[151] C. W. Mills, ""Ideal Theory" as Ideology," *Hypatia*, vol. 20, no. 3, pp. 165–184, 2005, Publisher: [Hypatia, Inc., Wiley].

[152] R. Kurzweil, *The singularity is near: when humans transcend biology*. New York, NY: Penguin Books, 2006, ISBN: 978-0-14-303788-0.

[153] M. Shanahan, *The Technological Singularity*. The MIT Press, Aug. 2015, ISBN: 978-0-262-33182-1.

[154] M. Sloane, E. Moss, O. Awomolo, and L. Forlano, "Participation Is not a Design Fix for Machine Learning," in *Equity and Access in Algorithms, Mechanisms, and Optimization*, Arlington VA USA: ACM, Oct. 2022, pp. 1–6, ISBN: 978-1-4503-9477-2.

[155] S. C. Slota, A. S. Hoffman, D. Ribes, and G. C. Bowker, "Prospecting (in) the data sciences," *Big Data & Society*, vol. 7, no. 1, p. 205 395 172 090 684, Jan. 2020.

[156] T. M. P. Reader, *The Myth of Objective Data*, Apr. 2023.

[157] M. Khan and A. Hanna, *The Subjects and Stages of AI Dataset Development: A Framework for Dataset Accountability*, SSRN Scholarly Paper, Rochester, NY, Sep. 2022.

[158] D. A. MacKenzie and J. Wajcman, Eds., *The social shaping of technology*, 2. ed., reprinted. Maidenhead: Open University Press, 2011, ISBN: 978-0-335-19914-3 978-0-335-19913-6.

[159] V. Kostakis, L. Lemos, and A. Kouvara, "Another Scalability is Possible! From Non-Scalability to Cosmolocal Scalability," *tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society*, vol. 22, no. 2, pp. 620–629, Nov. 2024.

[160] A. Schismenos, V. Niaros, and L. Lemos, "Cosmolocalism: Understanding the Transitional Dynamics Towards Post-Capitalism," *tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society*, pp. 670–684, Sep. 2020.

[161] G. C. Bowker and S. L. Star, *Sorting things out: classification and its consequences* (Inside technology), 1. paperback ed., 8. print. Cambridge, Mass.: MIT Press, 2008, ISBN: 978-0-262-52295-3 978-0-262-02461-7.

[162] N. Wiener, "Cybernetics," *Bulletin of the American Academy of Arts and Sciences*, vol. 3, no. 7, pp. 2–4, 1950, Publisher: American Academy of Arts & Sciences.

[163] S. Beer and S. Beer, *Decision and control: the meaning of operational research and management cybernetics* (The Stafford Beer classic library). Chichester ; New York: J. Wiley, 1994, ISBN: 978-0-471-06210-3 978-0-471-94838-4.

[164] Z. Thorwirth, *AI Winter: The Highs and Lows of Artificial Intelligence*, Sep. 2021.

[165] F. Heylighen and C. Joslyn, "Cybernetics and Second-Order Cybernetics," in *Encyclopedia of Physical Science and Technology*, Elsevier, 2003, pp. 155–169, ISBN: 978-0-12-227410-7.

[166] V. Kenny, ""There's Nothing Like the Real Thing". Revisiting the Need for a Third-Order Cybernetics," *Constructivist Foundations*, vol. 4, no. 2, pp. 100–111, 2009, Number: 2 Publisher: Alexander Riegler.

[167] J. Mökander, "Auditing of AI: Legal, Ethical and Technical Approaches," *Digital Society*, vol. 2, no. 3, p. 49, Nov. 2023.

[168] D. Haraway, "A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century," in *The Transgender Studies Reader*, Num Pages: 16, Routledge, 2006, ISBN: 978-0-203-95505-5.

[169] K. Knorr-Cetina, *Epistemic cultures: how the sciences make knowledge*. Cambridge, Mass: Harvard University Press, 1999, ISBN: 978-0-674-25893-8 978-0-674-25894-5.

[170] J. Desai, D. Watson, V. Wang, M. Taddeo, and L. Floridi, "The epistemological foundations of data science: A critical review," *Synthese*, vol. 200, no. 6, p. 469, Nov. 2022.

[171] J. McCarthy, "EPISTEMOLOGICAL PROBLEMS OF ARTIFICIAL INTELLIGENCE," in *Readings in Artificial Intelligence*, Elsevier, 1981, pp. 459–465, ISBN: 978-0-934613-03-3.

[172] P. Borsook, *Cyberselfish: a critical romp through the terribly libertarian culture of high tech*, 1st ed. New York: PublicAffairs, 2000, OCLC: 43481694, ISBN: 978-1-891620-78-2.

[173] N. C. M. Hartsock, "Standpoint Theories for the Next Century," *Women & Politics*, Jan. 1998, Publisher: Taylor & Francis Group.

[174] D. Haraway, *Manifestly Haraway* (Posthumanities 37). Minneapolis: University of Minnesota Press, 2016, ISBN: 978-0-8166-5047-7 978-0-8166-5048-4.

[175] T. H. Davenport and D. J. Patil, "Data Scientist: The Sexiest Job of the 21st Century," *Harvard Business Review*, Oct. 2012, Section: Analytics and data science.

[176] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, p. 4, 1965.

[177] K. Cukier and V. Mayer-Schönberger, "The Rise of Big Data: How It's Changing the Way We Think about the World," in *The Best Writing on Mathematics 2014*, M. Pitici, Ed., Princeton University Press, Dec. 2014, pp. 20–32, ISBN: 978-1-4008-6530-7.

[178] B. Cassel and H. Topi, "Strengthening data science education through collaboration," Arlington, VA, Workshop report, Jul. 2016.

[179] W. S. Cleveland, "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics," *International Statistical Review / Revue Internationale de Statistique*, vol. 69, no. 1, p. 21, Apr. 2001.

[180] U. Garczarek and D. Steuer, "Starting a Debate: Data Science – Occupation or Profession? A Discussion Paper," 2020, Publisher: Karlsruhe.

[181] C. Ley and S. P. A. Bordas, "What makes Data Science different? A discussion involving Statistics2.0 and Computational Sciences," *International Journal of Data Science and Analytics*, vol. 6, no. 3, pp. 167–175, Nov. 2018.

[182] C. Wu, *Statistics = Data Science?* University of Michigan, 1997.

[183] P. Naur, *Concise survey of computer methods*. Lund: Studentlitteratur, 1974, ISBN: 978-91-44-07881-6.

[184] J. J. Cuadrado-Gallego and Y. Demchenko, Eds., *The Data Science Framework: A View from the EDISON Project*. Cham: Springer International Publishing, 2020, ISBN: 978-3-030-51022-0 978-3-030-51023-7.

[185] W. Pietsch, *On the epistemology of data science: conceptual tools for a new inductivism* (Philosophical studies series Volume 148). Cham: Springer, 2022, ISBN: 978-3-030-86442-2 978-3-030-86441-5.

[186] J. v. Dijck, "Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology," *Surveillance & Society*, vol. 12, no. 2, pp. 197–208, May 2014.

[187] Q. E. McCallum, *Bad Data Handbook*. "O'Reilly Media, Inc.", Nov. 2012, Google-Books-ID: J6jTaDSg3sYC, ISBN: 978-1-4493-2188-8.

[188] T. Gillespie, *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*. New Haven: Yale University Press, 2018, OCLC: on1005113962, ISBN: 978-0-300-17313-0.

[189] J. C. YORK, *SILICON VALUES: the future of free speech under surveillance capitalism.* S.l.: VERSO BOOKS, 2022, OCLC: 1252050737, ISBN: 978-1-78873-881-1.

[190] M. Castillo, "Zuckerberg tells Congress Facebook is not a media company: 'I consider us to be a technology company'," *CNBC*, Apr. 2018.

[191] D. Baer, "Mark Zuckerberg Explains Why Facebook Doesn't 'Move Fast And Break Things' Anymore," *Business Insider*, May 2014.

[192] J. Jackson, M. Townsend, and L. Kassa, "Facebook 'lets vigilantes in Ethiopia incite ethnic killing'," *The Observer (The Guardian)*, Feb. 2022.

[193] E. Mackintosh, "Facebook knew it was being used to incite violence in Ethiopia. It did little to stop the spread, documents show," *CNN*, Oct. 2021.

[194] B. Ortutay, "Facebook's system approved dehumanizing hate speech," *PBS NewsHour*, Jun. 2022, Section: World.

[195] H. R. W. a. A. International, *"We will erase you from this land": crimes against humanity and ethnic cleansing in Ethiopia's Western Tigray Zone*. New York, N.Y.: Human Rights Watch ; Amnesty International, 2022, OCLC: 1339099667, ISBN: 978-1-62313-975-9.

[196] C. Allen, *Facebook's Content Moderation Failures in Ethiopia*, Apr. 2022.

[197] R. Bhatnagar, "In Ethiopia, Facebook allowed posts inciting violence to go viral for years. The company's response is both dismissive and ineffective.," *Media Matters for America*, Nov. 2021.

[198] Z. Zelalem and P. Guest, "Why Facebook keeps failing in Ethiopia," *Rest of World*, Nov. 2021.

[199] M. O. Board, *Announcing the Board's next cases*, Sep. 2021.

[200] M. O. Board, *Case decision 2021-014-FB-UA (FB-MP4ZC4CC)*, Dec. 2021.

[201] M. Miceli, T. Yang, L. Naudts, M. Schuessler, D. Serbanescu, and A. Hanna, "Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Trans-*

parency, Virtual Event Canada: ACM, Mar. 2021, pp. 161–172, ISBN: 978-1-4503-8309-7.

[202] R. Benjamin, "The Emperor's New Genes: Science, Public Policy, and the Allure of Objectivity," *The ANNALS of the American Academy of Political and Social Science*, vol. 661, no. 1, pp. 130–142, Sep. 2015.

[203] A. S. Chan, "Decolonial Computing and Networking Beyond Digital Universalism," *Catalyst: Feminism, Theory, Technoscience*, vol. 4, no. 2, pp. 1–5, Oct. 2018.

[204] S. Milan and E. Treré, "Big Data from the South(s): Beyond Data Universalism," *Television & New Media*, vol. 20, no. 4, pp. 319–335, May 2019.

[205] D. Coleman, "Digital Colonialism: The 21st Century Scramble for Africa through the Extraction and Control of User Data and the Limitations of Data Protection Laws," *Michigan Journal of Law and Race*, vol. 24, no. 417, 2019.

[206] B. Fiore-Gartland and G. Neff, "Communication, Mediation, and the Expectations of Data: Data Valences Across Health and Wellness Communities," *International Journal of Communication*, May 2015.

[207] S. Passi and P. Sengers, "Making data science systems work," *Big Data & Society*, vol. 7, no. 2, p. 205 395 172 093 960, Jul. 2020.

[208] S. Passi and S. Jackson, "Data Vision: Learning to See Through Algorithmic Abstraction," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, Portland Oregon USA: ACM, Feb. 2017, pp. 2436–2447, ISBN: 978-1-4503-4335-0.

[209] T. Gross, "Whistleblower Explains How Cambridge Analytica Helped Fuel U.S. 'Insurgency'," *NPR*, Oct. 2019.

[210] T. Olesen, "The Politics of Whistleblowing in Digitalized Societies," *Politics & Society*, vol. 47, no. 2, pp. 277–297, Jun. 2019, Publisher: SAGE Publications Inc.

[211] H. Von Unger, "Ethische Reflexivität in der ForschungspraxisEthical Reflexivity as Research Practice," *Historical Social Research Vol. 46*, vol. No. 2, Volumes per year: 1 , 2021, Publisher: GESIS - Leibniz-Institut für Sozialwissenschaften Version Number: 1.

[212] P. E. Agre, "The practical logic of computer work," in *Computationalism: New Directions*, M. Scheutz, Ed., MIT Press, 2002.

[213] L. A. Suchman and R. H. Trigg, "Artificial intelligence as craftwork," in *Understanding Practice: Perspectives on Activity and Context*, ser. Learning in Doing:

Social, Cognitive and Computational Perspectives, J. Lave and S. Chaiklin, Eds., Cambridge: Cambridge University Press, 1993, pp. 144–178, ISBN: 978-0-521-55851-8.

[214] S. Woolgar and G. Cooper, "Do Artefacts Have Ambivalence: Moses' Bridges, Winner's Bridges and other Urban Legends in S&TS," *Social Studies of Science*, vol. 29, no. 3, pp. 433–449, Jun. 1999.

[215] S. A. Cambo and D. Gergle, "Model Positionality and Computational Reflexivity: Promoting Reflexivity in Data Science," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI '22, New York, NY, USA: Association for Computing Machinery, Apr. 2022, pp. 1–19, ISBN: 978-1-4503-9157-3.

[216] B. Fish and L. Stark, "Reflexive Design for Fairness and Other Human Values in Formal Models," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '21, New York, NY, USA: Association for Computing Machinery, Jul. 2021, pp. 89–99, ISBN: 978-1-4503-8473-5.

[217] S. D. Hirsbrunner, M. Tebbe, and C. Müller-Birn, "From critical technical practice to reflexive data science," *Convergence*, p. 13 548 565 221 132 243, Nov. 2022, Publisher: SAGE Publications Ltd.

[218] E. Medina, "Designing Freedom, Regulating a Nation: Socialist Cybernetics in Allende's Chile," *Journal of Latin American Studies*, vol. 38, no. 3, pp. 571–606, Aug. 2006.

[219] S. Gerovitch, *From newspeak to cyberspeak: a history of Soviet cybernetics*. Cambridge, Mass: MIT Press, 2002, ISBN: 978-0-262-07232-8.

[220] AMPÈRE, ANDRÉ-MARIE, "Essai sur la philosophie des sciences ou exposition analytique d'une classification naturelle de toutes connaissances humaines," *Quai des Augustins*, vol. 55, 1834.

[221] D. Nutt and D. Railton, "The Sims: Real Life as Genre," *Information, Communication & Society*, vol. 6, no. 4, pp. 577–592, Dec. 2003, Publisher: Routledge _eprint: https://doi.org/10.1080/1369118032000163268.

[222] R. Vallee, "History of Cybernetics," *Systems Science and Cybernetics*, vol. 3, no. History of Cybernetics, p. 12, 2011.

[223] H. R. Maturana and F. J. Varela, *Autopoiesis and cognition: the realization of the living* (Boston studies in the philosophy of science v. 42). Dordrecht, Holland ; Boston: D. Reidel Pub. Co, 1980, ISBN: 978-90-277-1015-4 978-90-277-1016-1.

[224] *HOMOMORPHISM*.

[225] H. G. Frankfurt, *On bullshit*. Princeton, NJ: Princeton University Press, 2005, ISBN: 978-0-691-12294-6.

[226] M. T. Hicks, J. Humphries, and J. Slater, "ChatGPT is bullshit," *Ethics and Information Technology*, vol. 26, no. 2, p. 38, Jun. 2024.

[227] S. Turkle, *The second self: computers and the human spirit*, 20. anniversary ed., 1. MIT Press ed. Cambridge, Mass.: MIT Press, 2005, ISBN: 978-0-262-70111-2.

[228] A. P. Sage, "Cybernetics and Complex Adaptive Systems," in *Encyclopedia of Operations Research and Management Science*, S. I. Gass and M. C. Fu, Eds., Boston, MA: Springer US, 2013, pp. 339–348, ISBN: 978-1-4419-1153-7.

[229] W. R. Ashby, *An introduction to cybernetics*. Mansfield Centre, CT: Martino Publishing, 2015, OCLC: 1310736751.

[230] J. Kramer, "Is abstraction the key to computing?" *Communications of the ACM*, vol. 50, no. 4, pp. 36–42, Apr. 2007.

[231] G. Bowker, "How to be Universal: Some Cybernetic Strategies, 1943-70," *Social Studies of Science*, vol. 23, no. 1, pp. 107–127, Feb. 1993, Publisher: SAGE Publications Ltd.

[232] M. Feinberg, "A Design Perspective on Data," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Denver Colorado USA: ACM, May 2017, pp. 2952–2963, ISBN: 978-1-4503-4655-9.

[233] R. H. Godden and J. Hsy, "Universal Design and Its Discontents," in *Disrupting the Digital Humanities*, D. KIM and J. STOMMEL, Eds., Punctum Books, 2018, pp. 91–116, ISBN: 978-1-947447-71-4.

[234] C. Ryhl, "Simulating Disability in Universal Design Teaching: A Critique - Abandoning Try-It-Yourself in Teaching Universal Design in Architecture," *Studies in Health Technology and Informatics*, vol. 256, pp. 572–578, 2018.

[235] A. Hamraie, "Universal Design and the Problem of "Post-Disability" Ideology," *Design and Culture*, vol. 8, no. 3, pp. 285–309, Sep. 2016.

[236] S. Sweidan and K. Ejercito, "Non-user," *Internet Policy Review*, vol. 11, no. 2, Apr. 2022.

[237] J. Weizenbaum, *Computer power and human reason: from judgment to calculation*. San Francisco: Freeman, 1976, ISBN: 978-0-7167-0464-5 978-0-7167-0463-8.

[238] N. Ensmenger, ""Beards, Sandals, and Other Signs of Rugged Individualism": Masculine Culture within the Computing Professions," *Osiris*, vol. 30, no. 1, pp. 38–65, 2015, Publisher: [Saint Catherines Press, The University of Chicago Press, The History of Science Society].

[239] S. Brand, "SpaceWar: Fanatic Life and Symbolic Death Among the Computer Bums," *Rolling Stone*, Dec. 1972.

[240] J. Markoff, "What the Silicon Valley Prophet Sees on the Horizon," *The New York Times*, Mar. 2022.

[241] F. Turner, "Where the Counterculture Met the New Economy: The WELL and the Origins of Virtual Community," *Technology and Culture*, vol. 46, no. 3, pp. 485–512, 2005.

[242] J. Markoff, *Whole Earth: the many lives of Stewart Brand*. New York: Penguin Press, 2022, ISBN: 978-0-7352-2395-0.

[243] F. Turner, *From counterculture to cyberculture: Stewart Brand, the Whole Earth Network, and the rise of digital utopianism*. Chicago, Ill: University of Chicago Press, 2008, ISBN: 978-0-226-81742-2 978-0-226-81741-5.

[244] J. Markoff, *What the dormouse said: how the sixties counterculture shaped the personal computer industry*. New York: Penguin, 2006, ISBN: 978-0-14-303676-0.

[245] S. Brand, "Editor's Letter," *Whole Earth Catalog: Access to Tools*, 1968.

[246] W. Isaacson, *The innovators: how a group of hackers, geniuses, and geeks created the digital revolution*, First Simon & Schuster hardcover edition. New York: Simon & Schuster, 2014, ISBN: 978-1-4767-0869-0 978-1-4767-0870-6.

[247] R. Barbrook and A. Cameron, "The Californian Ideology," *Mute*, vol. 1, no. 3, Sep. 1995, Publisher: Mute Publishing Limited.

[248] M. G. Ames, *The charisma machine: the life, death, and legacy of One Laptop Per Child* (Infrastructures). Cambridge, MA: MIT Press, 2019, ISBN: 978-0-262-53744-5.

[249] M. Guzdial and J. Zhang, "Teaching CS humbly, and watching the AI revolution," *Communications of the ACM*, vol. 63, no. 5, pp. 24–25, Apr. 2020.

[250] L. Roberts, "The arpanet and computer networks," in *A history of personal workstations*, A. Goldberg, Ed., New York, NY, USA: ACM, Jan. 1988, pp. 141–172, ISBN: 978-0-201-11259-7.

[251] K. Flamm, *Creating the computer: government, industry, and high technology*. Washington, D.C: Brookings Institution, 1988, ISBN: 978-0-8157-2850-4 978-0-8157-2849-8.

[252] J. P. Barlow, *A Declaration of the Independence of Cyberspace*, Jan. 2016.

[253] L. Lessig, *Code*, Version 2.0. New York: Basic Books, 2006, OCLC: ocm77638613, ISBN: 978-0-465-03914-2.

[254] A. L. Bomse, "The Dependence of Cyberspace," *Duke Law Journal*, vol. 50, no. 6, p. 1717, Apr. 2001.

[255] J. P. Barlow, "Jackboots on the Infobahn," *Wired*, Apr. 1994, Section: tags.

[256] S. Ahmed, "This other and other others," *Economy and Society*, Nov. 2002, Publisher: Taylor & Francis Ltd.

[257] S. Brand, *Whole earth discipline: an ecopragmatist manifesto*. New York: Viking, 2009, ISBN: 978-0-670-02121-5.

[258] C. O'Neil, *Weapons of math destruction: how big data increases inequality and threatens democracy*, First edition. New York: Crown, 2016, ISBN: 978-0-553-41881-1 978-0-553-41883-5.

[259] J. A. English-Lueck, *Cultures@SiliconValley*, Second edition. Stanford, California: Stanford University Press, 2017, ISBN: 978-1-5036-0023-2 978-1-5036-0292-2.

[260] K. Schaffer, "The contested zone: Cybernetics, feminism and representation," *Journal of Australian Studies*, vol. 20, no. 50-51, pp. 157–164, Jan. 1996, Publisher: Routledge _eprint: https://doi.org/10.1080/14443059609387288.

[261] N. K. Hayles, "Technosymbiosis: Figuring (Out) Our Relations to AI," in *Feminist AI*, J. Browne, S. Cave, E. Drage, and K. McInerney, Eds., 1st ed., Oxford University PressOxford, Oct. 2023, pp. 1–18, ISBN: 978-0-19-288989-8 978-0-19-198282-8.

[262] S. Amrute, "Of Techno-Ethics and Techno-Affects," *Feminist Review*, vol. 123, no. 1, pp. 56–73, Nov. 2019.

[263] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer, "Wrangler: Interactive visual specification of data transformation scripts," in *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*, Vancouver, BC, Canada: ACM Press, 2011, p. 3363, ISBN: 978-1-4503-0228-9.

[264] Y. Mao *et al.*, "How Data ScientistsWork Together With Domain Experts in Scientific Collaborations: To Find The Right Answer Or To Ask The Right Question?" *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. GROUP, pp. 1–23, Dec. 2019.

[265] L. Cao, "Data Science: A Comprehensive Overview," *ACM Computing Surveys*, vol. 50, no. 3, pp. 1–42, Oct. 2017.

[266] G. Marchionini, "Information Science Roles in the Emerging Field of Data Science," *Journal of Data and Information Science*, vol. 1, no. 2, pp. 1–6, Sep. 2017.

[267] C. Bopp, E. Harmon, and A. Voida, "Disempowered by Data: Nonprofits, Social Enterprises, and the Consequences of Data-Driven Work," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Denver Colorado USA: ACM, May 2017, pp. 3608–3619, ISBN: 978-1-4503-4655-9.

[268] K. H. Pine *et al.*, "Data Work in Healthcare: Challenges for Patients, Clinicians and Administrators," in *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, Jersey City NJ USA: ACM, Oct. 2018, pp. 433–439, ISBN: 978-1-4503-6018-0.

[269] J. Lave and E. Wenger, *Situated learning: legitimate peripheral participation* (Learning in doing). Cambridge [England] ; New York: Cambridge University Press, 1991, ISBN: 978-0-521-41308-4 978-0-521-42374-8.

[270] M. Muller *et al.*, "Interrogating Data Science," in *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, Virtual Event USA: ACM, Oct. 2020, pp. 467–473, ISBN: 978-1-4503-8059-1.

[271] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: Mapping the debate," *Big Data & Society*, vol. 3, no. 2, p. 205 395 171 667 967, Dec. 2016.

[272] M. Feinberg, "The value of discernment: Making use of interpretive flexibility in metadata generation and aggregation," *Information Research-an International Electronic Journal*, vol. 22, no. 1, p. 22,

[273] G. C. Bowker, *Memory practices in the sciences* (Inside technology). Cambridge, Mass: MIT Press, 2005, OCLC: ocm60776866, ISBN: 978-0-262-02589-8.

[274] L. Gitelman, Ed., *"Raw data" is an oxymoron* (Infrastructures series). Cambridge, Massachusetts ; London, England: The MIT Press, 2013, ISBN: 978-0-262-51828-4.

[275] M. K. Scheuerman, A. Hanna, and E. Denton, "Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–37, Oct. 2021.

[276] T. Monroe-White, "Emancipatory Data Science: A Liberatory Framework for Mitigating Data Harms and Fostering Social Transformation," in *Proceedings of the 2021 on Computers and People Research Conference*, Virtual Event Germany: ACM, Jun. 2021, pp. 23–30, ISBN: 978-1-4503-8406-3.

[277] M. Feinberg, D. Carter, J. Bullard, and A. Gursoy, "Translating Texture: Design as Integration," in *Proceedings of the 2017 Conference on Designing Interactive Systems*, Edinburgh United Kingdom: ACM, Jun. 2017, pp. 297–307, ISBN: 978-1-4503-4922-2.

[278] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, and J. D. Weisz, "Expanding Explainability: Towards Social Transparency in AI systems," *arXiv:2101.04719 [cs]*, Jan. 2021, arXiv: 2101.04719.

[279] C. Sinders, R. Unicorn, C. Diehm, and I. A. Fumat, *Technically Responsible Knowledge (TKS)*, 2020.

[280] A. Meng, "Investigating the Roots of Open Data's Social Impact," *JeDEM - eJournal of eDemocracy and Open Government*, vol. 6, no. 1, pp. 1–13, Oct. 2014.

[281] A. Meng, C. DiSalvo, L. Tsui, and M. Best, "The social impact of open government data in Hong Kong: Umbrella Movement protests and adversarial politics," *The Information Society*, vol. 35, no. 4, pp. 216–228, Aug. 2019.

[282] L. Irani and J. Marx, *Redacted*. Taller California, 2021.

[283] K. Boehner and C. DiSalvo, "Data, Design and Civics: An Exploratory Study of Civic Tech," 2016, p. 12.

[284] M. Feinberg, "Reading databases: Slow information interactions beyond the retrieval paradigm," *Journal of Documentation*, vol. 73, no. 2, pp. 336–356, Mar. 2017.

[285] L. Koesten, K. Gregory, P. Groth, and E. Simperl, "Talking datasets – Understanding data sensemaking behaviours," *International Journal of Human-Computer Studies*, vol. 146, p. 102 562, Feb. 2021.

[286] A. Jain *et al.*, "Overview and Importance of Data Quality for Machine Learning Tasks," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Virtual Event CA USA: ACM, Aug. 2020, pp. 3561–3562, ISBN: 978-1-4503-7998-4.

[287] E. M. Bender and B. Friedman, "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 587–604, Dec. 2018.

[288] T. Gebru *et al.*, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, Dec. 2021.

[289] S. L. Bryant, A. Forte, and A. Bruckman, "Becoming Wikipedian: Transformation of participation in a collaborative online encyclopedia," in *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work - GROUP '05*, Sanibel Island, Florida, USA: ACM Press, 2005, p. 1, ISBN: 978-1-59593-223-5.

[290] O. St-Cyr, C. M. MacDonald, E. F. Churchill, J. J. Preece, and A. Bowser, "Developing a community of practice to support global HCI education," in *Extended abstracts of the 2018 CHI conference on human factors in computing systems*, ser. CHI EA '18, Number of pages: 7 Place: Montreal QC, Canada, New York, NY, USA: Association for Computing Machinery, 2018, pp. 1–7, ISBN: 978-1-4503-5621-3.

[291] S. Dasgupta and B. M. Hill, "Scratch Community Blocks: Supporting Children as Data Scientists," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Denver Colorado USA: ACM, May 2017, pp. 3620–3631, ISBN: 978-1-4503-4655-9.

[292] S. Hautea, S. Dasgupta, and B. M. Hill, "Youth Perspectives on Critical Data Literacies," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Denver Colorado USA: ACM, May 2017, pp. 919–930, ISBN: 978-1-4503-4655-9.

[293] L. Hardy, C. Dixon, and S. Hsi, "From Data Collectors to Data Producers: Shifting Students' Relationship to Data," *Journal of the Learning Sciences*, vol. 29, no. 1, pp. 104–126, Jan. 2020.

[294] M. H. Wilkerson *et al.*, "Data Transformations: Restructuring Data for Inquiry in a Simulation and Data Analysis Environment," p. 2, 2018.

[295] S. T. Roberts, *Behind the screen*. Yale University Press, 2019.

[296] C. Kiene *et al.*, "Volunteer Work: Mapping the Future of Moderation Research," in *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, Austin TX USA: ACM, Nov. 2019, pp. 492–497, ISBN: 978-1-4503-6692-2.

[297] M. Feinberg, W. Sutherland, S. B. Nelson, M. H. Jarrahi, and A. Rajasekar, "The New Reality of Reproducibility: The Role of Data Work in Scientific Research," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW1, pp. 1–22, May 2020.

[298] A. Crisan, B. Fiore-Gartland, and M. Tory, "Passing the Data Baton : A Retrospective Analysis on Data Science Work and Workers," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1860–1870, Feb. 2021.

[299] P. Pereira, J. Cunha, and J. P. Fernandes, "On Understanding Data Scientists," in *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, Dunedin, New Zealand: IEEE, Aug. 2020, pp. 1–5, ISBN: 978-1-72816-901-9.

[300] A. Y. Wang, A. Mittal, C. Brooks, and S. Oney, "How Data Scientists Use Computational Notebooks for Real-Time Collaboration," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–30, Nov. 2019.

[301] A. X. Zhang, M. Muller, and D. Wang, "How do Data Science Workers Collaborate? Roles, Workflows, and Tools," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW1, pp. 1–23, May 2020.

[302] L. F. Luna-Reyes, "The search for the data scientist: Creating value from data," *ACM SIGCAS Computers and Society*, vol. 47, no. 4, pp. 12–16, Jul. 2018.

[303] Y. Demchenko, L. Comminiello, and G. Reali, "Designing Customisable Data Science Curriculum Using Ontology for Data Science Competences and Body of Knowledge," in *Proceedings of the 2019 International Conference on Big Data and Education - ICBDE'19*, London, United Kingdom: ACM Press, 2019, pp. 124–128, ISBN: 978-1-4503-6186-6.

[304] D. Wang, J. Andres, J. Weisz, E. Oduor, and C. Dugan, "AutoDS: Towards Human-Centered Automation of Data Science," *arXiv:2101.05273 [cs]*, Jan. 2021, arXiv: 2101.05273.

[305] D. Wang *et al.*, "Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–24, Nov. 2019.

[306] J. Drozdal *et al.*, "Trust in AutoML: Exploring information needs for establishing trust in automated machine learning systems," in *Proceedings of the 25th International Conference on Intelligent User Interfaces*, Cagliari Italy: ACM, Mar. 2020, pp. 297–307, ISBN: 978-1-4503-7118-6.

[307] A. L. Strauss and J. M. Corbin, Eds., *Grounded theory in practice*. Thousand Oaks: Sage Publications, 1997, ISBN: 978-0-7619-0747-3 978-0-7619-0748-0.

[308] L. M. Given, Ed., *The Sage encyclopedia of qualitative research methods*. Los Angeles, Calif: Sage Publications, 2008, ISBN: 978-1-4129-4163-1.

[309] R. Edwards and J. Holland, *What is qualitative interviewing?* (What is? Research methods series). London : New Delhi: Bloomsbury, 2013, OCLC: ocn855705441, ISBN: 978-1-78093-852-3 978-1-84966-809-5.

[310] B. Smart, K. Peggs, and J. D. Burridge, Eds., *Observation methods* (SAGE benchmarks in social research methods). Los Angeles: SAGE, 2013, OCLC: ocn816163690, ISBN: 978-1-4462-0811-3.

[311] U. Flick, *An introduction to qualitative research*. 2009, OCLC: 1052103480, ISBN: 978-1-4462-4131-8 978-1-4462-0562-4 978-1-4462-8773-6.

[312] V. Braun, V. Clarke, N. Hayfield, and G. Terry, "Thematic Analysis," in *Handbook of Research Methods in Health Social Sciences*, P. Liamputtong, Ed., Singapore: Springer Singapore, 2019, pp. 843–860, ISBN: 978-981-10-5251-4.

[313] U. D. of Labor Statistics, *Labor Force Statistics from the Current Population Survey*, Jan. 2021.

[314] M. Muller and T. Erickson, "In the Data Kitchen: A Review (a design fiction on data science)," in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal QC Canada: ACM, Apr. 2018, pp. 1–10, ISBN: 978-1-4503-5621-3.

[315] N. Perlroth, *This is how they tell me the world ends: the cyberweapons arms race*. New York: Bloomsbury Publishing, 2021, ISBN: 978-1-63557-605-4.

[316] D. Ribes, "Notes on the Concept of Data Interoperability: Cases from an Ecology of AIDS Research Infrastructures," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, Portland Oregon USA: ACM, Feb. 2017, pp. 1514–1526, ISBN: 978-1-4503-4335-0.

[317] S. Cazacu, N. B. Hansen, and B. Schouten, "Empowerment approaches in digital civics," in *32nd australian conference on human-computer interaction*, ser. OzCHI '20, Number of pages: 8 Place: Sydney, NSW, Australia, New York, NY, USA: Association for Computing Machinery, 2020, pp. 692–699, ISBN: 978-1-4503-8975-4.

[318] S. Erete and J. O. Burrell, "Empowered participation: How citizens use technology in local governance," in *Proceedings of the 2017 CHI conference on human factors in computing systems*, ser. CHI '17, Number of pages: 13 Place: Denver, Col-

orado, USA, New York, NY, USA: Association for Computing Machinery, 2017, pp. 2307–2319, ISBN: 978-1-4503-4655-9.

[319] A. Knutas, V. Palacin, G. Maccani, and M. Helfert, "Software engineering in civic tech a case study about code for ireland," in *Proceedings of the 41st international conference on software engineering: Software engineering in society*, ser. ICSE-SEIS '19, Place: Montreal, Quebec, Canada Number of pages: 10, IEEE Press, 2019, pp. 41–50.

[320] J. Vertesi, "Seamful Spaces: Heterogeneous Infrastructures in Interaction," *Science, Technology, & Human Values*, vol. 39, no. 2, pp. 264–284, Mar. 2014.

[321] S. Inman and D. Ribes, ""Beautiful Seams": Strategic revelations and concealments," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, ser. CHI '19, Number of pages: 14 Place: Glasgow, Scotland Uk, New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–14, ISBN: 978-1-4503-5970-2.

[322] D. Dailey and K. Starbird, "Social media seamsters: Stitching platforms & audiences into local crisis infrastructure," in *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, ser. CSCW '17, Number of pages: 13 Place: Portland, Oregon, USA, New York, NY, USA: Association for Computing Machinery, 2017, pp. 1277–1289, ISBN: 978-1-4503-4335-0.

[323] N. Spadafora and T. Downes, "Scaffolding in Learning," in *Encyclopedia of Evolutionary Psychological Science*, T. K. Shackelford and V. A. Weekes-Shackelford, Eds., Cham: Springer International Publishing, 2020, pp. 1–4, ISBN: 978-3-319-16999-6.

[324] I. F. Ogbonnaya-Ogburu, A. D. Smith, A. To, and K. Toyama, "Critical Race Theory for HCI," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA: ACM, Apr. 2020, pp. 1–16, ISBN: 978-1-4503-6708-0.

[325] R. Delgado and J. Stefancic, *Critical race theory: An introduction*. NyU press, 2023, vol. 87.

[326] S. Costanza-Chock, *Design justice: community-led practices to build the worlds we need* (Information policy). Cambridge, Massachusetts: The MIT Press, 2020, ISBN: 978-0-262-04345-8.

[327] K. Levy, K. E. Chasalow, and S. Riley, "Algorithms and Decision-Making in the Public Sector," *Annual Review of Law and Social Science*, vol. 17, no. 1, annurev–lawsocsci–041221–023808, Oct. 2021.

[328] P. Freire, *Pedagogy of the oppressed*. Herder & Herder, 1970.

[329] P. Garcia *et al.*, "No: Critical refusal as feminist data practice," in *conference companion publication of the 2020 on computer supported cooperative work and social computing*, 2020, pp. 199–202.

[330] P. Garcia, T. Sutherland, N. Salehi, M. Cifor, and A. Singh, "No! Re-imagining data practices through the lens of critical refusal," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–20, 2022, Publisher: ACM New York, NY, USA.

[331] M. Cifor *et al.*, *Feminist data manifest-no*, 2019.

[332] S. Robinson, N. J. Bidwell, L. Maye, N. Pantidi, and C. Linehan, "Participation through substituting and refusing," in *Proceedings of the 16th Participatory Design Conference 2020 - Participation(s) Otherwise - Volume 2*, Manizales Colombia: ACM, Jun. 2020, pp. 143–147, ISBN: 978-1-4503-7606-8.

[333] Y. Milner, "Abolish big data," *University of California, Irvine*, vol. 8, 2019.

[334] S. Ahmed, "CONCLUSION 2. A killjoy manifesto," in *Living a feminist life*, Duke University Press, 2016, pp. 251–268.

[335] C. Rigg and K. Trehan, "Critical reflection in the workplace: Is it just too difficult?" *Journal of European Industrial Training*, 2008, Publisher: Emerald Group Publishing Limited.

[336] R. J. Cotter, "Reflexive spaces of appearance: Rethinking critical reflection in the workplace," *Human Resource Development International*, vol. 17, no. 4, pp. 459–474, Aug. 2014.

[337] C. Mouffe, "Deliberative democracy or agonistic pluralism?" *Social research*, pp. 745–758, 1999, Publisher: JSTOR.

[338] C. Mouffe, *Agonistics: Thinking the world politically*. Verso Books, 2013.

[339] B. Honig, "Political theory and the displacement of politics," in *Political theory and the displacement of politics*, Cornell University Press, 2016.

[340] H. Kraff, "A critical exploration of agonistic participatory design," *The Design Journal*, vol. 23, no. 1, pp. 31–48, 2020, Publisher: Taylor & Francis.

[341] E. Björgvinsson, P. Ehn, and P.-A. Hillgren, "Agonistic participatory design: Working with marginalised social movements," *CoDesign : international journal of cocre-*

ation in design and the arts, vol. 8, no. 2-3, pp. 127–144, 2012, Publisher: Taylor & Francis.

[342] N. Sawhney and A.-T. Tran, "Ecologies of contestation in participatory design," in *Proceedings of the 16th participatory design conference 2020-Participation (s) otherwise-volume 1*, 2020, pp. 172–181.

[343] A. A. Geppert and L. E. Forlano, "Design for equivalence: Agonism for collective emancipation in participatory design," in *Proceedings of the participatory design conference 2022-Volume 1*, 2022, pp. 158–168.

[344] R. Trigg and K. Ishimaru, "Integrating Participatory Design into everyday work at the Global Fund for Women," p. 22,

[345] J. Gartner and I. Wagner, "Mapping actors and agendas: Political frameworks of systems design and participation," *Human–Computer Interaction*, vol. 11, no. 3, pp. 187–214, 1996, Publisher: Taylor & Francis.

[346] E. E. Beck, "P for political: Participation is not enough," *Scandinavian Journal of Information Systems*, vol. 14, no. 1, p. 1, 2002, Publisher: Association for Information Systems.

[347] L. Huybrechts, H. Benesch, and J. Geib, "Institutioning: Participatory design, co-design and the public realm," *CoDesign : international journal of cocreation in design and the arts*, vol. 13, no. 3, pp. 148–159, 2017, Publisher: Taylor & Francis.

[348] M. Teli, M. Foth, M. Sciannamblo, I. Anastasiu, and P. Lyle, "Tales of institutioning and commoning: Participatory design processes with a strategic and tactical perspective," in *Proceedings of the 16th participatory design conference 2020-Participation (s) otherwise-volume 1*, 2020, pp. 159–171.

[349] Y. A. Rankin, J. O. Thomas, and N. M. Joseph, "Intersectionality in HCI: Lost in translation," *Interactions*, vol. 27, no. 5, pp. 68–71, Sep. 2020.

[350] I. Action, "Accomplices not allies: Abolishing the ally industrial complex," *Indigenous Action Media*, vol. 2, p. 2, 2014.

[351] G. Daswani, "On the whiteness of academia," *Everyday Orientalism*, 2020.

[352] S. Ahmed, "A phenomenology of whiteness," *Feminist theory*, vol. 8, no. 2, pp. 149–168, 2007, Publisher: Sage Publications Sage UK: London, England.

[353] A. Johnson, "An academic witness: White supremacy within and beyond academia," *The fire now: Anti-racist scholarship in times of explicit racial violence*, pp. 15–25, 2018, Publisher: Zed Books London, UK.

[354] R. Crooks and M. Currie, "Numbers will not save us: Agonistic data practices," *The Information Society*, pp. 1–19, May 2021.

[355] M. Liboiron, "Compromised agency: The case of BabyLegs," *Engaging Science, Technology, and Society*, vol. 3, pp. 499–527, 2017.

[356] E. Giaccardi, N. Cila, C. Speed, and M. Caldwell, "Thing Ethnography: Doing Design Research with Non-Humans," in *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, ser. DIS '16, New York, NY, USA: Association for Computing Machinery, Jun. 2016, pp. 377–387, ISBN: 978-1-4503-4031-1.

[357] A. McCosker, X. Yao, K. Albury, A. Maddox, J. Farmer, and J. Stoyanovich, "Developing data capability with non-profit organisations using participatory methods," *Big Data & Society*, vol. 9, no. 1, p. 20 539 517 221 099 882, Jan. 2022, Publisher: SAGE Publications Ltd.

[358] A. Voida, "Shapeshifters in the voluntary sector: Exploring the human-centered-computing challenges of nonprofit organizations," *Interactions*, vol. 18, no. 6, pp. 27–31, Nov. 2011.

[359] K. Nault, U. Ruhi, and O. Livvarcin, "Exploring the Applications & Challenges of Data Analytics in Non-Profit Organizations," *AMCIS 2020*, no. Session 8,

[360] S.-I. Wu and J.-M. Hung, "A performance evaluation model of CRM on non-profit organisations," *Total Quality Management & Business Excellence*, vol. 19, no. 4, pp. 321–342, Apr. 2008, Publisher: Routledge.

[361] S. Erete, E. Ryou, G. Smith, K. M. Fassett, and S. Duda, "Storytelling with Data: Examining the Use of Data by Non-Profit Organizations," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, ser. CSCW '16, New York, NY, USA: Association for Computing Machinery, Feb. 2016, pp. 1273–1283, ISBN: 978-1-4503-3592-8.

[362] E. Harmon, C. Bopp, and A. Voida, "The Design Fictions of Philanthropic IT: Stuck Between an Imperfect Present and an Impossible Future," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17, New York, NY, USA: Association for Computing Machinery, May 2017, pp. 7015–7028, ISBN: 978-1-4503-4655-9.

[363] S. J. Shapiro and V. Oystrick, "Three Steps Toward Sustainability: Spreadsheets as a Data-Analysis System for Non-Profit Organizations," *Canadian Journal of Program Evaluation*, vol. 33, no. 2, pp. 247–257, Sep. 2018, Publisher: University of Toronto Press.

[364] A. Wolff, D. Gooch, J. J. C. Montaner, U. Rashid, and G. Kortuem, "Creating an Understanding of Data Literacy for a Data-driven Society," *The Journal of Community Informatics*, vol. 12, no. 3, Aug. 2016, Number: 3.

[365] A. Rothschild, A. Meng, C. DiSalvo, B. Johnson, B. R. Shapiro, and B. DiSalvo, "Interrogating Data Work as a Community of Practice," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–28, Nov. 2022.

[366] S. Lee and P. Dourish, "Reconfiguring Data Relations: Institutional Dynamics around Data in Local Governance," *Proc. ACM Hum.-Comput. Interact.*, vol. 8, no. CSCW2, 420:1–420:28, Nov. 2024.

[367] R. Kitchin and T. Lauriault, *Towards Critical Data Studies: Charting and Unpacking Data Assemblages and Their Work*, SSRN Scholarly Paper, Rochester, NY, Jul. 2014.

[368] J. Gray, C. Gerlitz, and L. Bounegru, "Data infrastructure literacy," *Big Data & Society*, vol. 5, no. 2, p. 2 053 951 718 786 316, Jul. 2018, Publisher: SAGE Publications Ltd.

[369] L. M. Benjamin, A. Voida, and C. Bopp, "Policy fields, data systems, and the performance of nonprofit human service organizations," *Human Service Organizations: Management, Leadership & Governance*, vol. 42, no. 2, pp. 185–204, Mar. 2018, Publisher: Routledge _eprint: https://doi.org/10.1080/23303131.2017.1422072.

[370] M. Mitchell *et al.*, *Measuring Data*, arXiv:2212.05129 [cs], Dec. 2022.

[371] M. Gorli, D. Nicolini, and G. Scaratti, "Reflexivity in practice: Tools and conditions for developing organizational authorship," *Human Relations*, vol. 68, no. 8, pp. 1347–1375, Aug. 2015, Publisher: SAGE Publications Ltd.

[372] S. Leonelli, "Locating ethics in data science: Responsibility and accountability in global and distributed knowledge production systems," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2083, p. 20 160 122, Dec. 2016.

[373] D. McQuillan, *Resisting AI: an anti-fascist approach to artificial intelligence*. Bristol, UK: Bristol University Press, 2022, OCLC: on1328026349, ISBN: 978-1-5292-1349-2 978-1-5292-1350-8.

[374] A. Voida, E. Harmon, and B. Al-Ani, "Homebrew databases: Complexities of everyday information management in nonprofit organizations," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11, New York, NY, USA: Association for Computing Machinery, May 2011, pp. 915–924, ISBN: 978-1-4503-0228-9.

[375] T. A. Grossman, V. Mehrotra, and Özlük, "Lessons from Mission-Critical Spreadsheets," *Communications of the Association for Information Systems*, vol. 20, no. 1, Dec. 2007.

[376] National Council of Nonprofits, *(Mis)Understanding Overhead*.

[377] A. Fotopoulou, "Conceptualising critical data literacies for civil society organisations: Agency, care, and social responsibility," *Information, Communication & Society*, vol. 24, no. 11, pp. 1640–1657, Aug. 2021, Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2020.1716041.

[378] H. West and G. Green, "Because Excel will Mind Me! The State of Constituent Data Management in Small Nonprofit Organizations," *AMCIS 2008 Proceedings*, Jan. 2008.

[379] P. J. Stappers and E. Giaccardi, "Research through Design," in *The Encyclopedia of Human-Computer Interaction*, 2nd, Interaction Design Foundation, Nov. 2024.

[380] P. E. Agre, "Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI," in *Social Science, Technical Systems, and Cooperative Work*, G. Bowker, G. Bowker, S. L. Star, L. Gasser, and W. Turner, Eds., 0th ed., Psychology Press, May 2014, ISBN: 978-1-317-77876-9.

[381] R. Wells, "Google Sheets Vs. Microsoft Excel—Which Is Better?" *Forbes*, May 2024, Section: Careers.

[382] Z. J. Kunicki, N. S. Zambrotta, M. C. Tate, A. R. Surrusco, M. M. Risi, and L. L. Harlow, "Keep Your Stats in the Cloud! Evaluating the Use of Google Sheets to Teach Quantitative Methods," *Journal of Statistics Education*, vol. 27, no. 3, pp. 188–197, Sep. 2019, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10691898.201

[383] A. Puussaar, I. G. Johnson, K. Montague, P. James, and P. Wright, "Making Open Data Work for Civic Advocacy," *Proc. ACM Hum.-Comput. Interact.*, vol. 2, no. CSCW, 143:1–143:20, Nov. 2018.

[384] T. Davies and M. Frank, "'There's no such thing as raw data'. Exploring the sociotechnical life of a government dataset," *Proceedings of the 5th Annual ACM Web Science Conference*, 2013.

[385] K. L. Boyd, "Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–27, Oct. 2021.

[386] H. Karen and J. Sandra, "Contextual Inquiry: A Participatory Technique for System Design," in *Participatory Design*, Num Pages: 34, CRC Press, 1993, ISBN: 978-0-203-74433-8.

[387] V. Braun and V. Clarke, *Thematic Analysis*. Nov. 2021, Item ID: _:n0 Publication Title: 978-1-4739-5324-6.

[388] "ARIANE 5 Failure - Full Report," European Space Agency, Tech. Rep., Jul. 1996.

[389] DHS, *Over 20 Technology and Critical Infrastructure Executives, Civil Rights Leaders, Academics, and Policymakers Join New DHS Artificial Intelligence Safety and Security Board to Advance AI's Responsible Development and Deployment | Homeland Security*, Apr. 2024.

[390] A. McMillan-Major, E. M. Bender, and B. Friedman, "Data Statements: From Technical Concept to Community Practice," *ACM Journal on Responsible Computing*, p. 3 594 737, May 2023.

[391] O. Papakyriakopoulos *et al.*, "Augmented Datasheets for Speech Datasets and Ethical Decision-Making," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '23, New York, NY, USA: Association for Computing Machinery, Jun. 2023, pp. 881–904, ISBN: 9798400701924.

[392] M. Hamilton, "Inferring Race and Ethnicity Demographics from U.S. Census Data: Testing the Feasibility for Use in State Court Disparity Analyses," National Center for State Courts, Tech. Rep., Oct. 2022.

[393] K. Fiscella and A. M. Fremont, "Use of Geocoding and Surname Analysis to Estimate Race and Ethnicity," *Health Services Research*, vol. 41, no. 4 Pt 1, pp. 1482–1500, Aug. 2006.

[394] P. R. Kalluri, W. Agnew, M. Cheng, K. Owens, L. Soldaini, and A. Birhane, *The Surveillance AI Pipeline*, arXiv:2309.15084 [cs], Oct. 2023.

[395] S. Pink, M. Berg, D. Lupton, and M. Ruckenstein, *Everyday Automation: Experiencing and Anticipating Emerging Technologies*, 1st ed. London: Routledge, May 2022, ISBN: 978-1-00-317088-4.

[396] M. S. Silberman, L. Irani, and J. Ross, "Ethics and tactics of professional crowdwork," *XRDS: Crossroads, The ACM Magazine for Students*, vol. 17, no. 2, pp. 39–43, Dec. 2010.

[397] T. Shelton, *Business models for the social mobile cloud: transform your business using social media, mobile Internet, and cloud computing*. Hoboken, N.J: John

Wiley & Sons, 2013, Medium: electronic resource, ISBN: 978-1-118-49420-2 978-1-118-49419-6 978-1-118-49421-9.

[398] N. Tavanapour and E. A. C. Bittner, "The Collaboration of Crowd Workers.," in *26th European Conference on Information Systems: Beyond Digitization - Facets of Socio-Technical Change, ECIS 2018, Portsmouth, UK, June 23-28, 2018*, 2018, p. 65.

[399] R. Yang and H. Sun, "A new simulation framework for crowd collaborations," *International Journal of Crowd Science*, vol. 5, no. 1, pp. 2–16, Apr. 2021.

[400] C. Hießl, "Labour law for TOS and HITs? reflections on the potential for applying 'labour law analogies' to crowdworkers, focusing on employee representation," *Work Organisation, Labour & Globalisation*, vol. 12, no. 2, p. 38, 2018.

[401] A. Norton, *Automation and inequality The changing world of work in the global South*. International Institute for Environment and Development (IIED), 2017, OCLC: 7851137447, ISBN: 978-1-78431-493-4.

[402] C. Callison-Burch, "Crowd-Workers: Aggregating Information Across Turkers to Help Them Find Higher Paying Work.," in *Proceedings of the Seconf AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2014, November 2-4, 2014, Pittsburgh, Pennsylvania, USA*, 2014.

[403] A. Sriraman, J. Bragg, and A. Kulkarni, "Worker-Owned Cooperative Models for Training Artificial Intelligence," in *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, Portland Oregon USA: ACM, Feb. 2017, pp. 311–314, ISBN: 978-1-4503-4688-7.

[404] N. Salehi *et al.*, "We Are Dynamo: Overcoming Stalling and Friction in Collective Action for Crowd Workers," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, Seoul Republic of Korea: ACM, Apr. 2015, pp. 1621–1630, ISBN: 978-1-4503-3145-6.

[405] J. T. Jacques and P. O. Kristensson, "Crowdworker Economics in the Gig Economy," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow Scotland Uk: ACM, May 2019, pp. 1–10, ISBN: 978-1-4503-5970-2.

[406] A. Alkhatib, "To Live in Their Utopia: Why Algorithmic Systems Create Absurd Outcomes," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, May 2021, pp. 1–9, ISBN: 978-1-4503-8096-6.

[407] E. Christoforou, P. Barlas, and J. Otterbacher, "It's About Time: A View of Crowdsourced Data Before and During the Pandemic," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, May 2021, pp. 1–14, ISBN: 978-1-4503-8096-6.

[408] M. Budde, A. Exler, T. Riedel, M. Beigl, and A. Schankin, "Lessons from failures in designing and conducting experimental studies: A brief anecdotal tutorial," in *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, Maui Hawaii: ACM, Sep. 2017, pp. 992–999, ISBN: 978-1-4503-5190-4.

[409] T. Draws, A. Rieger, O. Inel, U. Gadiraju, and N. Tintarev, "A Checklist to Combat Cognitive Biases in Crowdsourcing," *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 9, no. 1, pp. 48–59, Oct. 2021, Section: Full Archival Paperss.

[410] J. Ramírez *et al.*, "On the State of Reporting in Crowdsourcing Experiments and a Checklist to Aid Current Practices," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–34, Oct. 2021.

[411] F. Bentley, K. O'Neill, K. Quehl, and D. Lottridge, "Exploring the Quality, Efficiency, and Representative Nature of Responses Across Multiple Survey Panels," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA: ACM, Apr. 2020, pp. 1–12, ISBN: 978-1-4503-6708-0.

[412] J. Pater, A. Coupe, R. Pfafman, C. Phelan, T. Toscos, and M. Jacobs, "Standardizing Reporting of Participant Compensation in HCI: A Systematic Literature Review and Recommendations for the Field," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, May 2021, pp. 1–16, ISBN: 978-1-4503-8096-6.

[413] *Guidelines for Academic Requesters - Turkopticon*, Oct. 2014.

[414] G. d'Eon, J. Goh, K. Larson, and E. Law, "Paying Crowd Workers for Collaborative Work," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–24, Nov. 2019.

[415] T. T. D. T. Nguyen, T. Garncarz, F. Ng, L. A. Dabbish, and S. P. Dow, "Fruitful Feedback: Positive Affective Language and Source Anonymity Improve Critique Reception and Work Outcomes," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, Portland Oregon USA: ACM, Feb. 2017, pp. 1024–1034, ISBN: 978-1-4503-4335-0.

[416] B. V. Hanrahan, D. Martin, J. Willamowski, and J. M. Carroll, "Investigating the Amazon Mechanical Turk Market Through Tool Design," *Computer Supported Cooperative Work (CSCW)*, vol. 27, no. 3-6, pp. 1255–1274, Dec. 2018.

[417] E. Loepp and J. T. Kelly, "Distinction without a difference? An assessment of MTurk Worker types," *Research & Politics*, vol. 7, no. 1, p. 205 316 801 990 118, Jan. 2020.

[418] L. Findlater, J. Zhang, J. E. Froehlich, and K. Moffatt, "Differences in Crowdsourced vs. Lab-based Mobile and Desktop Input Performance Data," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Denver Colorado USA: ACM, May 2017, pp. 6813–6824, ISBN: 978-1-4503-4655-9.

[419] R. A. Varanasi, A. Vashistha, D. Siddarth, V. Seshadri, and K. Bali, "Feeling Proud, Feeling Embarrassed: Experiences of Low-income Women with CrowdWork," in *In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22).*, New York, NY, USA: Association for Computing Machinery, 2022.

[420] D. Archambault, H. Purchase, and T. Hoßfeld, Eds., *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments: Dagstuhl Seminar 15481, Dagstuhl Castle, Germany, November 22 – 27, 2015, Revised Contributions* (Lecture Notes in Computer Science). Cham: Springer International Publishing, 2017, vol. 10264, ISBN: 978-3-319-66434-7 978-3-319-66435-4.

[421] S. E. T. Union, *ETUC resolution on digitalisation towards fair digital work*.

[422] U. Flick, *Doing grounded theory* (The SAGE qualitative research kit / edited by Uwe Flick 8. volume), 2nd edition. Los Angeles London New Dehli Singapore Washington DC Melbourne: SAGE, 2018, ISBN: 978-1-4739-1200-7.

[423] W. N. Espeland and M. L. Stevens, "A Sociology of Quantification," *European Journal of Sociology / Archives Européennes de Sociologie*, vol. 49, no. 3, pp. 401–436, Dec. 2008, Publisher: Cambridge University Press.

[424] A. Z. Jacobs and H. Wallach, "Measurement and Fairness," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21, New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 375–385, ISBN: 978-1-4503-8309-7.

[425] P. R. Maund *et al.*, "What motivates the masses: Understanding why people contribute to conservation citizen science projects," *Biological Conservation*, vol. 246, p. 108 587, Jun. 2020.

[426] T. D. Barnes, V. D. Beall, and M. R. Holman, "Pink-Collar Representation and Budgetary Outcomes in US States," *Legislative Studies Quarterly*, vol. 46, no. 1, pp. 119–154, 2021, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/lsq.12286.

[427] M. Crain, "Gender and Union Organizing," *ILR Review*, vol. 47, no. 2, pp. 227–248, Jan. 1994, Publisher: SAGE Publications Inc.

[428] M. Sainato, "'War of attrition': Why union victories for US workers at Amazon have stalled," *The Guardian*, Apr. 2023.

[429] OSTP, *Blueprint for an AI Bill of Rights*, Oct. 2022.

[430] A. Njanja, "Meta sued by Ethiopians and Kenyan rights group for fueling Tigray War," *TechCrunch*, Dec. 2022.

[431] J. Posada, "Embedded reproduction in platform data work," *Information, Communication & Society*, pp. 1–19, Mar. 2022.

[432] A.-T. Tran, A. Boone, C. A. Le Dantec, and C. DiSalvo, "Careful Data Tinkering," *Proc. ACM Hum.-Comput. Interact.*, vol. 6, no. CSCW2, 431:1–431:29, Nov. 2022.

[433] A. Klimczak, M. Wenka, M. Ganzha, M. Paprzycki, and J. Mańdziuk, "Towards Frugal Artificial Intelligence: Exploring Neural Network Pruning and Binarization," in *International Symposium on Intelligent Informatics*, S. M. Thampi, J. Mukhopadhyay, M. Paprzycki, and K.-C. Li, Eds., Singapore: Springer Nature, 2023, pp. 13–27, ISBN: 978-981-19-8094-7.

[434] T. Ménissier, "What ethics for the frugality of artificial intelligence?" In *Premières journées sur la recherche en apprentissage frugal*, Grenoble, France: MIAI Grenoble, Nov. 2022.

[435] K. Ng, B. Drenon, T. Gerken, and M. Cieslak, "What is DeepSeek - and why is everyone talking about it?" *BBC News*, Feb. 2025.

[436] V. Galetić, D. Sheldon, and A. Nottle, "Conceptual Knowledge Modelling for Human-AI Teaming in Data-Frugal Industrial Environments," in *Conceptual Knowledge Structures*, I. P. Cabrera, S. Ferré, and S. Obiedkov, Eds., Cham: Springer Nature Switzerland, 2024, pp. 217–232, ISBN: 978-3-031-67868-4.

[437] J. Stein, E. Dwoskin, H. Natanson, and J. O'Connell, "In chaotic Washington blitz, Elon Musk's ultimate goal becomes clear," *Washington Post*, Feb. 2025.

[438] P. Dave, "Elon Musk's DOGE Is Working on a Custom Chatbot Called GSAi," *Wired*, Feb. 2025, Section: tags.

[439] H. Barber and C. Stringer, "Musk to replace sacked government workers with machines," *The Telegraph*, Feb. 2025.

[440] F. Gilardi, M. Alizadeh, and M. Kubli, *ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks*, arXiv:2303.15056 [cs], Mar. 2023.

[441] W. Agnew *et al.*, *The illusion of artificial inclusion*, arXiv:2401.08572 [cs], Feb. 2024.

[442] S. Wang, Y. Liu, Y. Xu, C. Zhu, and M. Zeng, "Want To Reduce Labeling Cost? GPT-3 Can Help," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4195–4205.

[443] A. Wang, J. Morgenstern, and J. P. Dickerson, *Large language models cannot replace human participants because they cannot portray identity groups*, arXiv:2402.01908 [cs], Feb. 2024.

[444] C. Stokel-Walker, *Can AI Replace Human Research Participants? These Scientists See Risks*, Mar. 2024.

[445] I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, and R. Anderson, *The Curse of Recursion: Training on Generated Data Makes Models Forget*, arXiv:2305.17493 [cs], Apr. 2024.

[446] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal, "AI models collapse when trained on recursively generated data," *Nature*, vol. 631, no. 8022, pp. 755–759, Jul. 2024, Publisher: Nature Publishing Group.

[447] S. Alemohammad *et al.*, *Self-Consuming Generative Models Go MAD*, arXiv:2307.01850 [cs], Jul. 2023.

[448] C. Li *et al.*, "LabelAId: Just-in-time AI Interventions for Improving Human Labeling Quality and Domain Knowledge in Crowdsourcing Systems," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, ser. CHI '24, New York, NY, USA: Association for Computing Machinery, May 2024, pp. 1–21, ISBN: 9798400703300.

[449] S. Alemohammad *et al.*, *Self-Consuming Generative Models Go MAD*, arXiv:2307.01850 [cs], Jul. 2023.

[450] K. Fontichiaro and J. A. Oehrli, "Why Data Literacy Matters," *Knowledge Quest*, vol. 44, no. 5, pp. 21–27, 2016, Publisher: American Association of School Librarians ERIC Number: EJ1099487.

[451] A. Kaun *et al.*, "IN THE SHADOWS OF THE DIGITAL ECONOMY: THE GHOST WORK OF INFRASTRUCTURAL LABOR," *AoIR Selected Papers of Internet Research*, Oct. 2020.