Hanlin Li

The Dimensions of Data Labor: A Road Map for Researchers, Activists, and Policymakers to Empower Data Producers

Hanlin Li
hanlinl@berkeley.edu
University of California, Berkeley
Berkeley, CA, USA

Nicholas Vincent
nickvincent@u.northwestern.edu
University of California, Davis
Davis, CA, USA

Stevie Chancellor
steviec@umn.edu
University of Minnesota
Minneapolis, MN, USA

Brent Hecht
bhecht@northwestern.edu
Northwestern University
Evanston, IL, USA

ABSTRACT

Many recent technological advances (e.g. ChatGPT and search engines) are possible only because of massive amounts of user-generated data produced through user interactions with computing systems or scraped from the web (e.g. behavior logs, user-generated content, and artwork). However, data producers have little say in what data is captured, how it is used, or who it benefits. Organizations with the ability to access and process this data, e.g. OpenAI and Google, possess immense power in shaping the technology landscape. By synthesizing related literature that reconceptualizes the production of data for computing as "data labor", we outline opportunities for researchers, policymakers, and activists to empower data producers in their relationship with tech companies, e.g advocating for transparency about data reuse, creating feedback channels between data producers and companies, and potentially developing mechanisms to share data's revenue more broadly. In doing so, we characterize data labor with six important dimensions - legibility, end-use awareness, collaboration requirement, openness, replaceability, and livelihood overlap - based on the parallels between data labor and various other types of labor in the computing literature.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**.

KEYWORDS

user-generated data, empowerment, data leverage

ACM Reference Format:
Hanlin Li, Nicholas Vincent, Stevie Chancellor, and Brent Hecht. 2023. The Dimensions of Data Labor: A Road Map for Researchers, Activists, and Policymakers to Empower Data Producers. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3593013.3594070

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06…$15.00
https://doi.org/10.1145/3593013.3594070

1 INTRODUCTION

Technology users generate large troves of data in their daily interactions with computing systems, e.g. behavior logs, content, and personal information. Currently, this data primarily benefits just a small set of technology organizations that are equipped with the means and resources to collect, process, and model data at scale for their own benefits (e.g. insights, models, sales of services and advertisements). For example, publicly available texts and artwork enabled the creation of generative AI models like ChatGPT and Dall-E because model developers were able to scrape and process data from billions of web pages [1]. Conversely, data producers like artists, writers, and users have little to no power in deciding how their data is used or who it benefits [4, 7, 42, 63]. This power imbalance between data producers and technology operators has manifested in public outcries about industry practices in the tech sector. For example, emerging generative AI models such as Stable Diffusion, Dall-E, and GitHub Copilot have sparked extensive criticism among artists and programmers because of these models' unapproved reuse of their work and implications on future employment opportunities [79–81]. More broadly, social media users have long protested the monetization of user data and the corporate surveillance practices that tend to go with it [45].

Given data producers' lack of power over the data they generate, researchers, policymakers, and activists have advocated for a new producer-oriented paradigm shift to increase the voice of the data-generating public – understanding data generation as a form of labor, or "data labor" [7]. Supporters of this approach have argued that treating data as an outcome of social labor instead of "exhaust" will pave the way for more broadly distributing the power and benefits of data [86], and scholars have addressed what this may look like in practice. Initial (yet abstract) proposals include supporting "data unions" [63] or "mediators of individual data" [42] that negotiate data use terms with technology firms on behalf of their data-producing "union" members [63], drafting legislation that would grant users greater control over the data they produce [1, 76], and creating tools to support user-driven collective action [20, 86].

[1] https://commoncrawl.org/2022/10/sep-oct-2022-crawl-archive-now-available/

Li, H., Vincent, N., Chancellor, S., & Hecht, B. (2023). The Dimensions of Data Labor: A Road Map for Researchers, Activists, and Policymakers to Empower Data Producers. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 1151–1161.