weizenbaum
DISCUSSION PAPER

June 2025
48

Tianling Yang, Christian Strippel, Alexandra Keiner, Dylan Baker, Alexis Chávez, Krystal Kauffman, Marc Pohl, Caroline Sinders, Milagros Miceli

Ethics of Data Work

Principles for Academic Data Work Requesters

ABOUT THE AUTHORS

Tianling Yang \\ Weizenbaum Institute \ Technische Universität Berlin
Christian Strippel \\ Weizenbaum Institute
Alexandra Keiner \\ Weizenbaum Institute
Dylan Baker \\ DAIR Institute
Alexis Chávez \\ Turkopticon
Krystal Kauffman \\ Turkopticon \ DAIR Institute
Marc Pohl \\ Independent Researcher
Caroline Sinders \\ Convocation Design + Research
Milagros Miceli \\ Weizenbaum Institute \ Technische Universität Berlin \ DAIR Institute

Contact: christian.strippel@weizenbaum-institute.de, m.miceli@tu-berlin.de

ABOUT THIS PAPER

This paper presents principles for academic data work requesters, developed as part of the short project "Ethics of Data Work" at the Weizenbaum Institute in Berlin, Germany. During two workshops held at the Institute in September and November of 2024, an interdisciplinary team of scholars, practitioners, and data workers discussed relevant issues in the context of data work requests and ways to build equitable and sustainable systems for both data workers and requesters. In our view, ethical principles for academic data requesters are central to such systems. The authors want to thank the Weizenbaum Institute for funding this project.

ABOUT THE WEIZENBAUM INSTITUTE

The Weizenbaum Institute is a joint project funded by the German Federal Ministry of Research, Technology and Space (BMFTR) and the State of Berlin. It conducts interdisciplinary and basic research on the digital transformation of society and provides evidence- and value-based options for action in order to shape digitalization in a sustainable, self-determined and responsible manner.

Weizenbaum Discussion Paper

Ethics of Data Work

Principles for Academic Data Work Requesters

Tianling Yang, Christian Strippel, Alexandra Keiner, Dylan Baker, Alexis Chávez, Krystal Kauffman, Marc Pohl, Caroline Sinders, Milagros Miceli

Abstract

The growing use of machine learning (ML) in academic research has led to a rising demand for large, labeled datasets. While the field initially relied on the labor of students and research assistants to label data, as models grew larger and more complex, there was a shift towards relying on large-scale, low-cost platforms like Amazon Mechanical Turk (MTurk) to label data at scale. However, this shift comes with serious ethical concerns. Now part of a massive industry, many data work companies exploit workers, leaving many workers facing low wages and precarious working conditions, with little institutional oversight or protection. Despite the centrality of this labor to modern research, ethical codes and guidelines from academic societies rarely address the implications of outsourcing data work to platform-based workers. This paper advocates for the development of research ethics standards that ensure fair and responsible collaboration with data workers. We begin by defining the concept of "data work" and assessing how current ethical frameworks address it. We then highlight ongoing initiatives aimed at improving ethical regulation. Based on two focus groups and two expert workshops held at the Weizenbaum Institute in 2024, we propose a set of principles for academic data work requesters to guide ethical engagement with platform-based workers. Finally, we outline future steps for integrating these principles into scientific ethical codes and day-to-day research practices.

Table of Contents

1 Introduction

With the increasing popularity of machine learning (ML) methods in scientific research, the need for large-scale manual data work, such as labeling, has grown significantly in recent years (Schmidt, 2022). For example, a supervised ML model that classifies the sentiment of an internet comment does so by “learning” from a training data set of comments labeled by humans. For small datasets, these kinds of labels could be obtained by researchers themselves. Increasingly, however, much larger datasets are used, necessitating the use of data work platforms like Amazon Mechanical Turk (MTurk) or Prolific to crowdsource this labor (Chan & Holosko, 2016). Now, even tasks that could have historically been carried out within research teams are often completed on such platforms, with several authors citing significantly reduced costs and greater accessibility as factors in this shift (Buhrmester et al., 2011; Silberman et al., 2018; Geiger et al., 2020).

In recent years, numerous studies have shown that many data workers suffer from precarious working conditions (e.g., Miceli et al., 2024). For this reason, we believe it is important to reflect on the use of such services within scientific research from an ethics perspective and to develop ethical standards for platform-based collaboration with data workers. Today, this ethical self-regulation work within academia is still in its infancy. As we will demonstrate, scientific associations largely lack ethical codes and guidelines around the use of data work for research.

In this paper, we propose a set of ethical principles for requesting platform-based data for academic research purposes, aiming for these standards to be adopted by scientific associations. First, we describe what we mean by “data work” (Section 2). Then, we review the codes of ethics of various scientific organizations to assess how well they already address data work (Section 3). Next, we list existing initiatives that aim to improve the ethical regulation of data work (Section 4). Finally, we introduce our ethical principles for academic data work requesters (Section 5). These principles were developed following two focus groups with data workers and requesters, as well as two expert workshops held at the Weizenbaum Institute in September and November of 2024. These workshops brought together an international and interdisciplinary team of scholars, practitioners, and data workers to discuss problems and needs based on practical examples and concrete scenarios. In the final chapter, we suggest future steps for implementing corresponding regulations in ethical codes and guidelines for scientific research (Section 6).

2 What is "Data Work"?

Data is essential to the development of artificial intelligence (AI). In particular, ML systems that involve larger classification or predictive systems require massive data to train models, which will identify patterns and leverage them to inform decisions and make predictions. Understood as "the labor involved in the collection, curation, classification, labeling, and verification of data" (Miceli & Posada, 2022), data work is key to making data available, readable, and analyzable to machines. It is important to highlight that, despite full-automation narratives, data work remains "a structural rather than a temporary input to AI production" (Tubaro et al., 2020, p. 4).

Data work has been rendered invisible in AI production. Investigating the processes of producing ML datasets reveals that data is carefully constructed, requiring the collaboration of multiple actors, such as reviewers, quality analysts, project managers, and clients, in addition to data workers (Miceli et al., 2020). Data work is typically outsourced through Business Process Outsourcing companies (BPOs) or crowdsourcing platforms to workers in the Global South, who are underpaid, work with little social protection, and under intense work control and surveillance (Miceli et al., 2022). This allows clients, who are often based in the Global North, to access data services at low prices, which further deepens economic disparities, reinforces power imbalances, and perpetuates patterns of extraction and exploitation (Tubaro & Casilli, 2019).

In academic contexts, data work is needed across various disciplines, such as bioscience (Davies & Holmer, 2024), social sciences (Plantin, 2019), and ocean science (Neang et al., 2020), and can vary in the level of professionalism and institutionalization. It is marked by different craft practices (Thomer et al., 2022), coordinative work, and collaboration strategies (Neang et al., 2020). Despite the importance of data work to science communities, workers often struggle with short-term contracts, lack of financial support, and unclear career pathways (Davies & Holmer, 2024). Their labor in data production often remains invisible, under-recognized, and uncredited (Plantin, 2019). Davies and Holmer (2024) further argue that such under-recognition shows the devaluation of data work as knowledge practice, resulting in epistemic injustice for workers. Moreover, data work in academia has been increasingly outsourced outside the academic spheres (see e.g., Geiger et al., 2020), such as through crowdsourcing platforms like MTurk, Prolific, and CloudResearch that enable researchers to obtain data quickly and inexpensively (Buhrmester et al., 2011) but raise concerns around data quality and research ethics. This is particularly relevant in the context of ML model development, as many teams span both academia and industry, and the demand for outsourcing data work is notably high.

Recently, researchers have advocated for incorporating labor issues into the understanding of the social impacts and harms of AI-based systems (Solaiman et al., 2024) and into ethical considerations (Miceli, 2022; Casilli, 2024; Muldoon et al., 2025). This paper joins this line of work and calls for ethical data work in scientific practice.

3 Lack of Ethical Self-Regulation in Academia

Although data work is widely used in academic research, its ethical implications are rarely discussed within scientific communities. As Geiger et al. (2020) pointed out, for example, ML application papers rarely specify how annotations were conducted or whether crowdsourcing platforms were used. Even when they do, the papers often fail to report how the workers were compensated. In addition, numerous ethical guidelines for the use of AI have been developed in recent years by universities, scientific institutions, and tech companies. However, these guidelines primarily address issues related to the deployment of AI, such as its impact on labor, data protection, privacy, and transparency (Jobin et al., 2019; Hagendorff, 2020). Nevertheless, the working conditions of data workers—those who contribute to the creation and maintenance of AI systems—are rarely, if ever, mentioned.

Scientific associations usually develop and maintain ethical guidelines that address standards for scientific research, including the use of AI technologies and data collection. However, a review of these frameworks reveals a profound dearth of explicit standards for data work. Only very few frameworks in academia explicitly mention ethical concerns about the use of platforms such as MTurk. For example, the “Ethical Guidelines 3.0”, issued by the Association of Internet Researchers (AoIR, 2019), addresses the poor, often task-based, compensation, which may affect the quality of the resulting data (p. 22). However, it does not provide concrete recommendations on how academic requesters should approach data work. Furthermore, the “Code of Ethics” of the Neural Information Processing Systems Conference (NeurIPS, 2023) includes a passage on “fair wages,” according to which academics who “make use of crowdsourcing or contract work for a particular task” as part of their research, “must respect the minimum hourly rate in the region where the work is carried out.”

Existing frameworks do address important ethical principles in academia, such as working conditions, responsibility, and authorship, which could be extended to data labor practices. For example, the “Code of Ethics and Professional Conduct” of the Association for Computing Machinery (ACM, 2018) emphasizes the respect and protection of privacy regarding the collection and use of data involving personal information. It also calls for high quality in not only the products but also the work processes, ensuring transparency in project communication. Moreover, it highlights the importance of improving the quality of working life, taking into consideration “the personal and professional development, accessibility requirements, physical safety, psychological well-being, and human dignity of all workers”. However, the ACM Code of Ethics does not establish specific guidelines to ensure the well-being of workers, nor does it mention data workers.

Similarly, the “Code of Ethics and Professional Conduct” issued by the Association for Information Systems (AIS, 2019) emphasizes responsibility in the use of information systems. While the guidelines remain more general, they state that the AIS aims to improve the quality of life, well-being, and working conditions of individuals. However, the term “individuals” remains ambiguous—it could refer to users, data subjects, developers, or researchers.

The Ethical Guidelines of the German Informatics Society (GI, 2018) emphasize the importance of communicating with affected communities to understand their needs and interests. They advocate for fair terms of employment, inclusive and supportive organizational structures, and encourage efforts to improve local and global living conditions. Moreover, they underscore the responsibility to "contribute to the betterment of local and global living conditions" throughout all stages of IT systems, from design to application. However, similarly to ACM, GI does not explicitly name data workers.

Other ethical frameworks highlight aspects that could also be extended to data work, such as the question of authorship. For example, the "Guidelines for Safeguarding Good Research Practice" of the German Research Foundation (DFG, 2022) emphasize that contributions such as "gathering, collection, acquisition or provision of data, software or sources, or the analysis/evaluation or interpretation of data, sources and conclusions drawn" (p. 18) should be acknowledged. Even when such contributions do not qualify for authorship, they should be documented in other ways, such as in footnotes.

The German Data Ethics Commission (2019) likewise points to data rights and argues that data rights holders should retain control over their data and may be entitled to "an economic share in profits derived with the help of the data" (p. 10). This principle could be extended to data workers, supporting the argument that their contributions to research projects should be adequately compensated.

4 Existing Research and Initiatives

As the exploitative nature of data work has become more widely reported, efforts have been made to define guidelines for more ethical data work practices.

The Fairwork project at the Oxford Internet Institute (OII) has published written principles that center fairness regarding payment, working conditions, terms of contract, organizational management, and workers' access to representation and collective bargaining (Fairwork, 2025). These principles are further differentiated between location-based platform work and cloudwork (online platform work), as the working conditions and risks workers are faced with can vary to a large extent. Fairwork, along with the Global Partnership on AI and the OECD, regularly publishes reports evaluating different platforms against their principles.

The worker-advocacy group Turkopticon (2025) has developed guidelines for requesters and for journalists and researchers who study workers on MTurk. Taking worker-centered and community-based approaches, they call for transparency in the information about requesters to ensure accountability and responsibility, the provision of proper time estimates, timely approval of completed tasks, respect for worker privacy, and strict adherence to MTurk's terms of service. Moreover, requesters and researchers should make explicit the conditions for rejecting work, avoid blocking workers to prevent duplicate participants, establish effective and responsive communication with workers, provide fair payment at or above the minimum Turking wage, clearly explain the requirements for additional rewards, and provide compensation for qualifier or screener surveys.

The Partnership on AI, a coalition that brings together tech companies, civil society organizations, and academic researchers, has shared resources to improve the conditions for what they refer to as "data enrichment workers" (PIA, 2025). Their "Data Enrichment Sourcing Guidelines" call for ensuring payment above the local living wage, designing and running a pilot before the start of projects, identifying suitable workers and workforce for projects to include diverse and representative perspectives, providing proper instructions and training for workers that are verified by workers, and setting up effective communication mechanisms between workers and practitioners (PAI, 2022a). In addition, Partnership on AI provides a checklist for good task instructions (PAI, 2022b). Moreover, their white paper "Responsible Sourcing of Data Enrichment Services" further advocates for establishing worker-minded quality assurance, seeking feedback from workers, and recognizing their contribution to larger projects (PAI, 2021). It should be noted that the Partnership on AI, a non-profit partnership of over 100 institutions, includes tech companies like Google and Meta, as well as data labeling BPOs like Prolific.

In a report by the International Labour Organization (ILO), Berg and colleagues (2018) outlined 18 criteria for fair data work, including but not limited to the recognition of workers' employment status, establishment of legal processes to communicate their needs and demands, the application of minimum wage to microtask workers, and transparency in payment and fees.

Germany's largest labor union, IG Metall, has been at the forefront of organizing for the rights of crowdworkers in Germany. Convening a working group of European and North American unions, labor confederations, and worker organizations, IG Metall and its collaborators published the 2016 "Frankfurt Declaration on Platform-Based Work," calling for a variety of worker protections. This work was preceded by the 2015 launch of their Fair Crowd Work Initiative website, a hub for crowdworkers resources run by IG Metall. Their website includes legal resources for workers in Germany and Austria, connections to several unions globally that advocate for crowdworkers, and platform reviews based on interviews with workers. At the same time, IG Metall and several crowdsourcing platforms collectively worked on a code of conduct in 2015 (Gegenhuber et al., 2022) and released "Principles for Paid Crowdsourcing/Crowdworking" in 2017 (IG Metall, 2017), highlighting issues such as workers' rights, fair payment, task description and management, constructive feedback and open communication. Gegenhuber and colleagues (2022) studied IG Metall's work and the emergence of one collective institutional agreement they made between their union and several crowdsourcing platforms. Drawing parallels to other forms of collective and union-inclusive governance models for highly decentralized work, such as in the garment industry, and to corporate social responsibility voluntary self-regulation frameworks, they identify the motivating forces behind the emergence of this labor ethics agreement. These include platforms' motivations to avoid regulation and the union's interest in creating working agreements that address new realities of digital work.

The proposals outlined above primarily address companies requesting data work and crowdsourcing platforms, aiming to promote ethical and responsible sourcing in industry pipelines. However, with the exception of Turkopicon's guidelines for academic researchers and the proposal for CrowdWorkSheets by Díaz et al. (2022), there is inadequate attention to the specific responsibility of academic requesters. We aim to fill this gap with a call for enhanced self-regulation and reflexivity within the academic community.

5 Principles for Academic Data Work Requesters

5.1 Preamble

Data is at the heart of many research disciplines. Researchers increasingly hire data workers from around the world to create, collect, classify, and clean data. However, their labor and contributions to research often remain invisible and unrecognized. This invisibilization is reinforced within many research ethical guidelines, most of which do not yet address data work. At the same time, data workers face many problems in the context of their research involvement that need to be accounted for. To address these issues, a team of scholars, practitioners, and data workers has developed the following ethical principles to guide future collaborations. We urge academic institutions to incorporate these principles into their policies and encourage researchers to use them as a starting point for their ethical reflection in a rapidly changing field.

5.2 General Principles

Researchers are responsible for the external data workers they engage to create, collect, or process data. This responsibility is not absolved by the involvement of third-party organizations, such as platforms or business process outsourcing companies (BPOs). Researchers should respect data workers as research collaborators and experts in their own work practices. They should recognize workers' labor and contributions to knowledge production and provide appropriate payment according to living wages, collective agreements, expertise, and task type. Researchers must do everything in their power to try to ensure the well-being, good working conditions, the privacy, and autonomy of data workers. This includes transparent and respectful communication, as well as careful consideration of workers' diverse contexts. Researchers must be aware of the uneven power relations that shape these aspects of short-term employment with data workers and be accountable for any extractive practices involved in this process. This also means refraining from conducting projects in which the well-being of data workers cannot be ensured.

5.3 Key Challenges and Recommendations

To clarify these ethical principles, we will examine four key challenges—collaboration, payment, working conditions, and communication—in greater detail and provide specific recommendations for addressing them.

5.3.1 Data Workers as Research Collaborators

When data work is outsourced to platforms or BPOs, a hierarchy emerges between requesters and data workers. Consequently, requesters often do not recognize the work as part of the research process, but rather as an external service. Terms such as "gig worker" or "clickworker" devalue the work by suggesting that it is a simple and comparatively marginal task performed by low-skilled individuals. This perspective is reflected in the payment, working conditions, and communication with data workers (see below). Recognizing data workers as research collaborators can address this hierarchical relationship, providing opportunities to better capture their expertise and improve research outcomes. To address this challenge, we recommend the following:

- Acknowledge the collaboration with data workers by granting them authorship or other forms of credit, such as a mention in the methodology, preface, footnotes, or acknowledgments.
- Listen to and reflect on criticism and suggestions from data workers.
- Respect work practices and avoid micromanagement and unnecessary complications, such as excessive time controls.
- Respect the expertise of data workers in the process of co-producing datasets.
- Be as transparent as possible with data workers about the context and goals of the project in order to mitigate alienation.
- Avoid using terms that devalue the expertise and contributions of data workers to the research project, such as "micro", "gig", "click", and "ghost" workers.

5.3.2 Appropriate Payment

Most data workers are subject to low wages or unpaid work. Underpayment can be exacerbated by workers' geographical location and the terms and conditions of platforms and BPOs. Since requesters usually do not pay data workers directly, they are often unaware of the additional costs associated with platform provision fees or payment provider transaction fees when calculating payments for their tasks. In addition, data workers' expertise is often not recognized, resulting in their categorization as low-skilled. This significantly impacts their remuneration by requesters, platforms, and BPOs (see above). Furthermore, low pay is sometimes justified by framing it as philanthropy, for example, by offering job opportunities to underserved populations. But requesting data work is not charity. The AI industry needs these workers just as much as these workers need jobs. Demanding data work from vulnerable people is less about philanthropy and more about the ability to pay low wages. To ensure fair wages and prevent the exploitation of data workers, we recommend the following:

- Respect living wages and collective agreements. If living wages are not in place, adhere to the standards of the country in which the requesting institution is located.
- Respect that data workers are experts who should be rewarded as such.

- Fairly estimate the time per task, taking into account the length of the description, a poor internet connection, and potential errors within the task.
- Avoid unpaid work due to potential rejection, errors, or unclear tasks.
- Take into account the cost of foreign currency exchange or the provision of the platform when calculating payment.
- Pay for extra work, such as identifying errors or improving the task.
- Make sure that the payment covers the necessary resources to complete the task, such as internet access or specialized software.
- Approach decisions such as rejection or blocking carefully, as they could negatively affect data workers' performance history, which could impact their future assignments or payments.
- Pay on time and do not assign further tasks or work until payment has been secured, even if research organizations sometimes need more time to process payments.
- Offer workers partial compensation for completing a certain percentage of the task successfully (e.g., 90 % or 95 %).
- When choosing a platform or BPO, consider their payment policies.

5.3.3 Working Conditions

Data workers may experience exhaustion due to an extensive workload, mental health issues resulting from exposure to disturbing content, unfair contracts with platforms or BPOs, instability in their work, and unpaid time spent accessing and preparing tasks, training, and waiting. They also bear the costs of technical issues with infrastructure and internet connectivity. Sometimes, they face concerns about personal safety resulting from excessive and identifiable data collection. In this context, to pursue ethical data work, we recommend the following:

- Recognize unions and worker representatives as credible authorities, on par with BPO managers and platforms.
- Read about the working conditions of the platforms or BPOs you are considering.
- Take appropriate measures and provide benefits for tasks involving disturbing content. This includes providing resources for psychological support and extra payment, extending deadlines, and allowing breaks during tasks.
- If the task requires the use of additional online tools, such as specific annotation tools, please ensure that workers are provided with the necessary access.

5.3.4 Transparency and Respectful Communication

Data workers interact with numerous data requesters every day. As a result, they encounter many conflicts, unanswered questions, and other communication issues. The anonymity of platform-based data work hides their personalities, circumstances, and contexts from requesters, which often causes requesters to make little or no effort to communicate. This can negatively affect the well-being of data workers. To promote ethical communication with data workers, we recommend the following:

- Communicate with data workers in a respectful and non-hierarchical manner. Keep in mind that they are not a homogeneous group. They come from different backgrounds, live in different contexts, and work under different circumstances.
- Create clear and concise task instructions. If possible, consult people outside of the research team to test the instructions before involving data workers.
- Be transparent about the purpose of the data work and how the data will be used, whether at the beginning or end of employment (e.g., debriefing).
- Provide trigger warnings for data work that includes disturbing content, or that may offend certain social groups.
- Respect workers' right to voice their feedback by providing them with contact information so they can seek clarification, offer suggestions, and raise disputes. Be available to answer questions, especially at the beginning of the project.
- Respect the privacy of data workers and limit the collection of their data to what is strictly necessary. While contact details can sometimes be necessary, excessive collection of personal data can undermine the protection of workers' privacy and even their personal safety.
- Approach potential conflicts with care, being aware of the power imbalances between researchers and data workers.

6 Conclusion and Future Perspectives

This paper outlines a set of ethical principles to guide academic researchers who outsource tasks to data workers. Based on two focus groups and two expert workshops held at the Weizenbaum Institute in 2024, the principles propose that academic requesters recognize data workers as collaborators rather than mere proxies, ensure fair compensation and working conditions, provide sufficient time to complete tasks, avoid harm from exposing workers to sensitive content, and maintain transparency and openness in communication. These principles respond to a notable gap in existing ethical frameworks: While industry best practices are increasingly discussed, academic institutions have largely overlooked the ethical responsibilities of researchers who rely on data work. By prompting reflection on the lived realities of data workers, we aim to foster more just and sustainable practices in academic ML research.

Because data production involves a global and heterogeneous workforce, there are no one-size-fits-all answers to the ethical concerns that arise. Power dynamics between requesters and data workers influence the issues raised in this paper and demand continued reflexivity. The lack of ethical guidance for academic requesters is particularly pressing given that many researchers, unlike their industry counterparts, have limited experience with data work and often do not know where to begin. In the focus group we conducted with academic researchers, it became clear that even well-intentioned academics struggle to translate ethical concerns into concrete practices. This paper addresses that gap by offering clear principles.

But principles alone are not enough. Ensuring ethical academic data work requires institutional support and practical tools. As next steps, we aim to collaborate with professional associations to push for the integration of these principles into their ethics codes. Additionally, we are developing a tool to help academic requesters design task instructions aligned with the principles presented in this paper, aiming to translate these principles into day-to-day research practices.

References

ACM – Association for Computing Machinery. (2018). ACM code of ethics and professional conduct. https://www.acm.org/code-of-ethics

AIS – Association for Information Systems. (2018). AIS code of ethics and professional conduct. https://aisnet.org/page/CodeOfEthics

Anonymous Workers. (2024). Mind over Moderation. In M. Miceli, A. Dinika, K. Kauffman, C. Salim Wagner, & L. Sachenbacher (eds.), *The data workers' inquiry*. https://data-workers.org/berlin

Berg, J., Furrer, M., Harmon, E., Rani, U., & Silberman, M. S. (with International Labour Organization). (2018). *Digital labour platforms and the future of work: Towards decent work in the online world*. International Labour Organization.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3–5. https://doi.org/10.1177/1745691610393980

Casilli, A. A. (2024). 'End-to-end' ethical AI. Taking into account the social and natural environments of automation. *In Artificial intelligence, labour and society* (pp. 83 – 92). ETUI.

Chan, C., & Holosko, M. J. (2016). An overview of the use of Mechanical Turk in behavioral sciences: Implications for social work. *Research on Social Work Practice, 26*(4), 441 – 448. https://doi.org/10.1177/1049731515594024

Davies, S. R., & Holmer, C. (2024). Care, collaboration, and service in academic data work: Biocuration as 'academia otherwise.' *Information, Communication & Society, 27*(4), 683 – 701. https://doi.org/10.1080/1369118X.2024.2315285

Data Ethics Commission. (2019). Opinion of the Data Ethics Commission. https://www.bfdi.bund.de/SharedDocs/Downloads/EN/Datenschutz/Data-Ethics-Commission_Opinion.pdf?__blob=publicationFile&v=1

DFG – Deutsche Forschungsgemeinschaft. (2022). Code of conduct: Guidelines for safeguarding good research practice. Zenodo. https://doi.org/10.5281/zenodo.6472827

Díaz, M., Kivlichan, I., Rosen, R., Baker, D., Amironesei, R., Prabhakaran, V., & Denton, E. (2022). CrowdWorkSheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2342 – 2351. https://doi.org/10.1145/3531146.3534647

Fairwork. (2025). Principles. https://fair.work/en/fw/principles/

Gegenhuber, T., Schuessler, E., Reischauer, G., & Thäter, L. (2022), Building collective institutional infrastructures for decent platform work: The development of a crowdwork agreement in Germany. In A. A. Gümüsay, E. Marti, H. Trittin-Ulbrich, & C. Wickert (eds.), *Organizing for societal grand challenges* (pp. 43–68). Emerald Publishing Limited.. https://doi.org/10.1108/S0733-558X20220000079004

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., & Huang, J. (2020). Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 325–336. https://doi.org/10.1145/3351095.3372862

GI – German Informatics Society. (2018). Ethical guidelines of the German Informatics Society. https://gi.de/ethicalguidelines

Hagendorff, T. (2020). *The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

IG Metall. (2017). Code of Conduct – Grundsätze für bezahltes Crowdsourcing / Crowdworking [Code of conduct – Principles for paid crowdsourcing / crowdworking]. https://www.igmetall.de/download/docs_Code_of_Conduct__7a063a6feada7aee2b3fb656764022c4c4c57302.pdf

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Miceli, M., & Posada, J. (2022). The data-production dispositif. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW2), 1 – 37. https://doi.org/10.1145/3555561

Miceli, M., Posada, J., & Yang, T. (2022). Studying up machine learning data: Why talk about bias when we mean power? *Proceedings of the ACM on Human-Computer Interaction, 6*(GROUP), 1 – 14. https://doi.org/10.1145/3492853

Muldoon, J., Cant, C., Graham, M., & Ustek Spilda, F. (2025). The poverty of ethical AI: Impact sourcing and AI supply chains. *AI & Society, 40*, 529 – 543. https://doi.org/10.1007/s00146-023-01824-9

Neang, A. B., Sutherland, W., Beach, M. W., & Lee, C. P. (2021). Data integration as coordination: The articulation of data work in an ocean science collaboration. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW3), 1 – 25. https://doi.org/10.1145/3432955

NeurIPS – Neural Information Processing Systems Conference. (2023). Code of ethics. https://neurips.cc/public/EthicsGuidelines

PAI – Partnership on AI. (2021). Responsible sourcing of data enrichment services. partnershiponai.org/responsible-sourcing

PAI – Partnership on AI. (2022a). Data enrichment sourcing guidelines. https://partnershiponai.org/wp-content/uploads/2022/11/data-enrichment-guidelines.pdf

PAI – Partnership on AI. (2022b). Good instructions checklist for data enrichment projects. https://partnershiponai.org/wp-content/uploads/2022/11/data-enrichment-checklist.pdf

PIA – Partnership on AI. (2025). Improving conditions for data enrichment workers. https://partnershiponai.org/responsible-sourcing-library/

Plantin, J.-C. (2019). Data cleaners for pristine datasets: Visibility and invisibility of data processors in social science. *Science, Technology, & Human Values, 44*(1), 52–73. https://doi.org/10.1177/0162243918781268

Schmidt, F. A. (2022). The planetary stacking order of multilayered crowd-AI systems. In M. Graham & F. Ferrari (eds.), *Digital work in the planetary market* (pp. 137 – 156). MIT Press. https://doi.org/10.7551/mitpress/13835.003.0012

Silberman, M. S., Tomlinson, B., LaPlante, R., Ross, J., Irani, L., & Zaldivar, A. (2018). Responsible research with crowds. *Communications of the ACM, 61*(3), 39 – 41. https://doi.org/10.1145/3180492

Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., … Subramonian, A. (2024). Evaluating the social impact of generative AI systems in systems and society [Preprint]. *arXiv*. http://arxiv.org/abs/2306.05949

Thomer, A. K., Akmon, D., York, J. J., Tyler, A. R. B., Polasek, F., Lafia, S., Hemphill, L., & Yakel, E. (2022). The craft and coordination of data curation: Complicating workflow views of data science. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW2), 1 – 29. https://doi.org/10.1145/3555139

Tubaro, P., & Casilli, A. A. (2019). Micro-work, artificial intelligence and the automotive industry. *Journal of Industrial and Business Economics, 46*(3), 333 – 345. https://doi.org/10.1007/s40812-019-00121-1

Tubaro, P., Casilli, A. A., & Coville, M. (2020). The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence. *Big Data & Society, 7*(1). https://doi.org/10.1177/2053951720919776

Turkopticon. (2025). Basics of how to be a good requester. https://blog.turkopticon.net/?page_id=212

\\ Imprint

Tianling Yang, Christian Strippel, Alexandra Keiner, Dylan Baker, Alexis Chávez, Krystal Kauffman, Marc Pohl, Caroline Sinders, Milagros Miceli

Ethics of Data Work. Principles for Academic Data Work Requesters

Weizenbaum Discussion Paper #48
June 2025

ISSN 2748-5587 \ DOI 10.34669/WI.DP/48

Weizenbaum-Institut e. V.
Hardenbergstraße 32 \ 10623 Berlin \ Tel.: +49 30 700141-001
info@weizenbaum-institut.de \ www.weizenbaum-institut.de

COORDINATION: Dr. Moritz Buchner \ **TYPESETTING:** Luisa Le van, M.A., Atelier Hurra

weizenbaum institut

With funding from the:

Federal Ministry of Research, Technology and Space

BERLIN
Senate Department for Higher Education and Research, Health and Long-Term Care