Commentary

BIG DATA & SOCIETY

Big Data & Society
April–June: 1–6

DOI: 10.1177/20539517251351320
journals.sagepub.com/home/bds
Sage

Better pay, clearer guidance: Investing in the working conditions of artificial intelligence data workers

Johann Laux[1,2], Fabian Stephany[1,2,3] and Alice Liefgreen[4]

Abstract

The production of artificial intelligence (AI) requires human labour, with tasks ranging from well-paid engineering work to often-outsourced data work. This commentary explores the economic and policy implications of improving working conditions for AI data workers, specifically focusing on the impact of clearer task instructions and increased pay for data annotators. It contrasts rule-based and standard-based approaches to task instructions, revealing evidence-based practices for increasing accuracy in annotation and lowering task difficulty for annotators. AI developers have an economic incentive to invest in these areas as better annotation can lead to higher quality AI systems. The findings have broader implications for AI policy beyond the fairness of labour standards in the AI economy. Testing the design of annotation instructions is crucial for the development of annotation standards as a prerequisite for scientific review and effective human oversight of AI systems in protection of ethical values and fundamental rights.

Keywords

Artificial intelligence, data annotation, human labour, working conditions, AI policy, experimental design

The economic case for improving working conditions in artificial intelligence data work

As artificial intelligence (AI) systems spread across sectors, demand for human labour in their development is rising. The economic value of AI skills such as machine learning is reflected in higher pay of up to 23% for workers with AI skills within the same occupation (Gonzalez Ehlinger and Stephany, 2023; Stephany and Teutloff, 2024). Besides well-paid engineering work, human labour is needed to prepare the data used to design and develop AI systems (Muldoon, Cant et al., 2024a; Miceli and Posada, 2022; Crawford, 2021; Tubaro et al., 2020). These 'AI data workers' collect, curate, annotate, and evaluate data sets and AI model outputs (Muldoon, Cant et al., 2024a: 1–2).

Take data annotation as an example. Annotation by humans, including labelling images and tagging texts, is essential to training, testing, and validating AI and machine learning models (Patra et al., 2023; Shemtob et al., 2023). It is the most time-consuming part of data preparation in the AI lifecycle (Muldoon, Cant et al., 2024a: 10), directly impacts the efficiency and accuracy of AI models (Rädsch et al., 2023), and can be as crucial to model performance as computational power (Eisenmann et al., 2023). Data annotation can both mitigate and introduce biases as well as ensure fairness and lead to discriminatory outcomes (Chen et al., 2023; Denton et al., 2021; Paullada et al., 2020; Yang et al., 2020). Labelling errors in commonly used test data sets such as *ImageNet* can have adverse effects, as AI practitioners often choose which model to deploy based on accuracy benchmarked against the test data set (Northcutt et al., 2021).

To reduce costs, developers aim to minimise the required amount of human-labelled data or to automate data annotation (Wang et al., 2023). Firms also outsource annotation work (Tubaro, 2021), both to outsourcing companies and online labour platforms (Muldoon, Cant et al., 2024a;

[1]Oxford Internet Institute, University of Oxford, Oxford, UK
[2]Humboldt Institute for Internet and Society, Berlin, Germany
[3]Bruegel, Brussels, Belgium
[4]Department of Language and Cognition, University College London, London, UK

Corresponding author:
Johann Laux, Oxford Internet Institute, University of Oxford, Oxford, UK; Humboldt Institute for Internet and Society, Berlin, Germany.
Email: johann.laux@oii.ox.ac.uk

Miceli and Posada, 2022; Ørting et al., 2020; Tubaro et al., 2020). The outsourced labour often goes to countries in the Global South, where wages are lower and labour regulations are less defined (Le Ludec et al., 2023; Braesemann et al., 2022; Graham and Anwar, 2019). This development has led to debates about fair compensation and work conditions for AI data workers (Muldoon, Graham et al., 2024b; Wong, 2023; Gray and Suri, 2019; Wood et al., 2019).

Presumably, AI developers have an incentive to invest in better working conditions. Improvements could lead to higher accuracy in annotation outcomes and, thus, higher quality data sets and AI applications trained on this data. Taking this economic view, how could working conditions be improved?

According to a recent survey, data annotators ('annotators' from now on) in biomedical image labelling stated that the greatest cause of problems in their daily work was unclear task instructions (Rädsch et al., 2023). Previous research has shown that how task instructions are formulated on online labour platforms impacts the quality of outcomes (Gillier et al., 2018). Besides clarifying instructions, working conditions could be improved through raising wages. Earlier research suggested that payment rates on online labour platforms have no detectable influence on data quality outcomes–until the platforms' workforce moved to the Global South, monetary compensation became its primary motivation, and compensation rates began to make a difference for data quality (Litman et al., 2015). The quality of instructions and the fairness of payment for AI data workers are thus significant factors for producing high-quality AI systems. Recent research has shown that both can be subject to a unifying economic analysis (Laux et al., 2023). This comment argues that clearer guidance and better pay for AI data workers should be important objectives for AI policy. Beyond the fairness of labour standards in the AI economy, the development of empirically tested annotation instruction standards is an important prerequisite for scientific review and effective human oversight of AI systems.

Organisational choices: task instructions as rules or standards

How should organisations best formulate instructions for data annotation? This question is surprisingly underexplored and public quality standards for data annotation are still lacking (Rädsch et al., 2023). A recent experiment in biomedical image labelling shows that increasing the 'information density' in instructions through showing exemplary images yields better outcomes, while solely extending text descriptions does not (Rädsch et al., 2023). In their randomised control trial with annotators in an image labelling task for accessible building design, Laux et al. (2023) draw on a distinction from the field of law and economics: the difference between formulating norms as rules or as standards (Korobkin, 2000).

Rules bind a decisionmaker to a determinate result if specific facts occur. Standards give the decisionmaker more discretion, as they require the application of a background principle or policy to a particularised set of facts (Korobkin, 2000). A 55 miles per hour speed limit is a rule; 'drive at reasonable speed' is a standard as it still requires the decisionmaker to determine which speed level is reasonable (Kaplow, 1992). Rules provide clarity and predictability in their application due to their upfront specificity, whereas standards are more vague and more adaptable to specific contexts (Casey and Niblett, 2017). Rules create higher up-front decision costs but are advantageous for frequent and homogenous behaviour as once formulated, they benefit from economies of scale. *Vice versa*, for infrequent and heterogeneous behaviour, standards are preferable, as providing specific details for future scenarios can be deferred until the application stage (Casey and Niblett, 2017; Kaplow and Shavell, 2002).

Experimental outcomes: rules and pay benefit accuracy

Laux et al. (2023) test the effects of rules versus standards with annotators tasked to label a data set of 100 images of building entrances as either barrier-free or non-barrier-free for a person in a wheelchair. Figure 1 shows an example trial with the 'rules' condition.

Laux et al. (2023) recruited 307 participants via an online experiment platform (Prolific). Participants were randomly assigned to three different task instructions: the 'rules' and 'standards' conditions as ideal-types and an 'incomplete rules' condition accounting for a possible lack of full information for the rule-maker. While the rules condition provides an extensive textual description of what constitutes a barrier-free entrance (see Figure 1), incomplete rules lack some of the information provided through rules (such as the need for a handrail), and standards merely state that entrances must be accessible for wheelchair users without help by another person. Similar to (Rädsch et al., 2023), there is thus variation in information density between conditions, albeit purely in text and not through adding or removing instructive images. Importantly, participants were also randomly assigned to two different payment conditions: while every participant was paid a baseline pay of £4.50, one group was informed about the possibility of receiving a performance-determined bonus of £3. Instruction design and monetary compensation were thus combined in one experimental setting, allowing to compare the cost-efficiency of both treatments. The study thus featured six different groups (3×2 design), which contained roughly 50 participants in each group.

Figure 2 shows the impact of the task instruction conditions and monetary incentives on the accuracy of annotators' work. Figure 2(A) displays the accuracy rates in the three different instructional conditions, highlighting that a rule-based

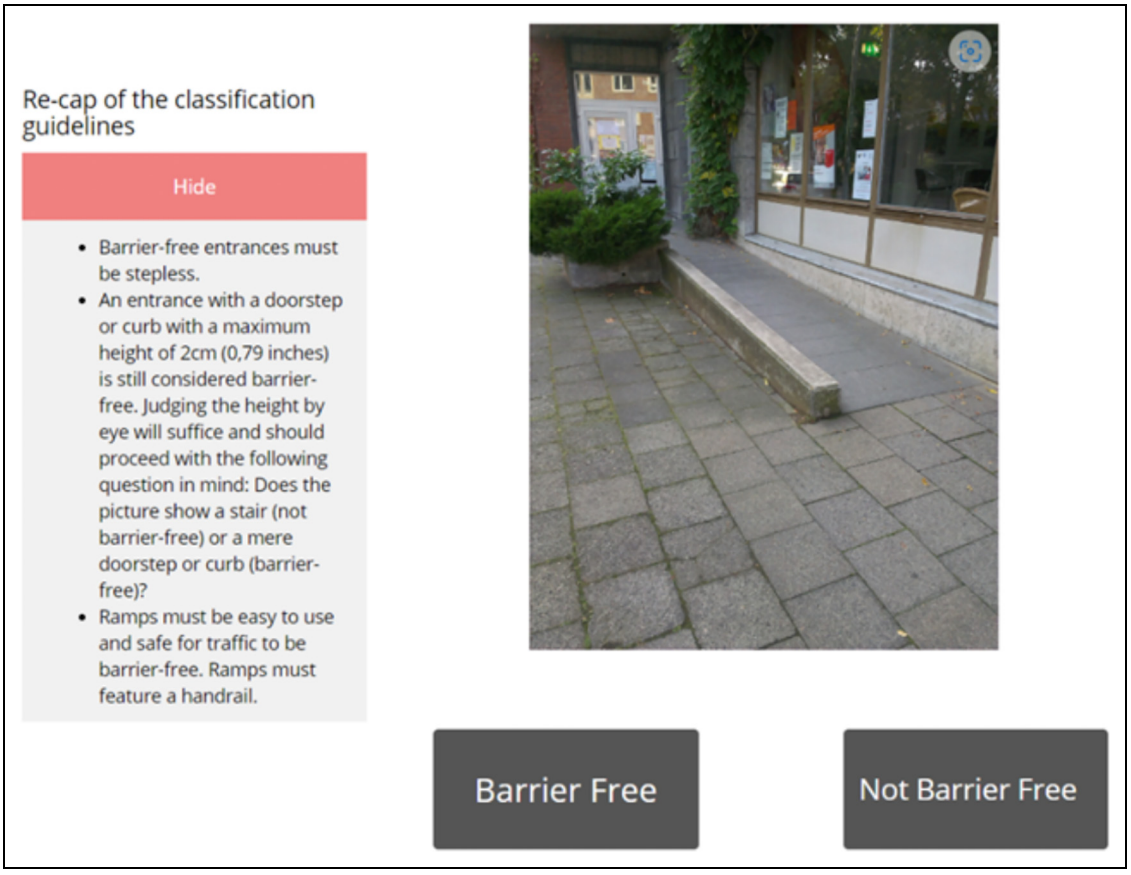

Figure 1. Example trial: image of an entrance together with the rules condition as shown to participants. Source: Laux et al. (2023).

instruction yields the highest accuracy at 85.6%. This represents an improvement of 11 percentage points over the accuracy observed under the standard-based condition, which stands at 74.9%. This significant improvement underscores the effectiveness of rule-based instructions in enhancing accuracy. In contrast, the effect of the monetary incentive on accuracy is much less pronounced. Accuracy increases modestly from 78.2% with no incentive to 80.1% with the incentive, reflecting a mere 1.9 percentage point improvement. Figure 2(C) provides a consolidated view of these findings, showing that the best-performing annotators were working both under rules *and* with the monetary incentive (86.7% accuracy).

Annotators' perceptions: rules are helpful and reduce difficulty

An important aspect of improving working conditions for AI data workers is how the workers themselves perceive the task instructions. Laux et al. (2023) survey annotators as to how helpful rules, incomplete rules, and standards are for their work. Figure 3 examines the results. Figure 3(A) shows that the number of images that annotators report having trouble classifying is notably lower when annotators are working under rules. Specifically, annotators reported an average of 14 difficult images under the rules condition, whereas they reported 18 difficult images under the standards condition. This difference corresponds to a reduction of approximately 22% in the number of difficult images (i.e. roughly four images) when rules are used, highlighting the effectiveness of clear, rule-based guidance in reducing task difficulty. Figure 3(B) further supports the advantage of rules by examining the perceived helpfulness of these instructions. Here, 57% of workers reported their rule-based instructions were very helpful, in contrast to only 44% of workers who found the standard-based instructions to be very helpful. This comparison indicates that rules not only reduce perceived task difficulty but also enhance the overall usefulness of the guidance provided to annotators. Regarding the additional category of incomplete rules, it is worth noticing that the missing information (such as the requirement of a handrail) rendered incomplete rules significantly less helpful than complete rules yet still led to fewer images reported to be difficult to annotate.

Implications for artificial intelligence policy

The results in Laux et al. (2023) suggest that AI developers have an economic incentive to invest in improved working conditions of annotators. By investing in rules and monetary incentives, they receive higher accuracy in labelling

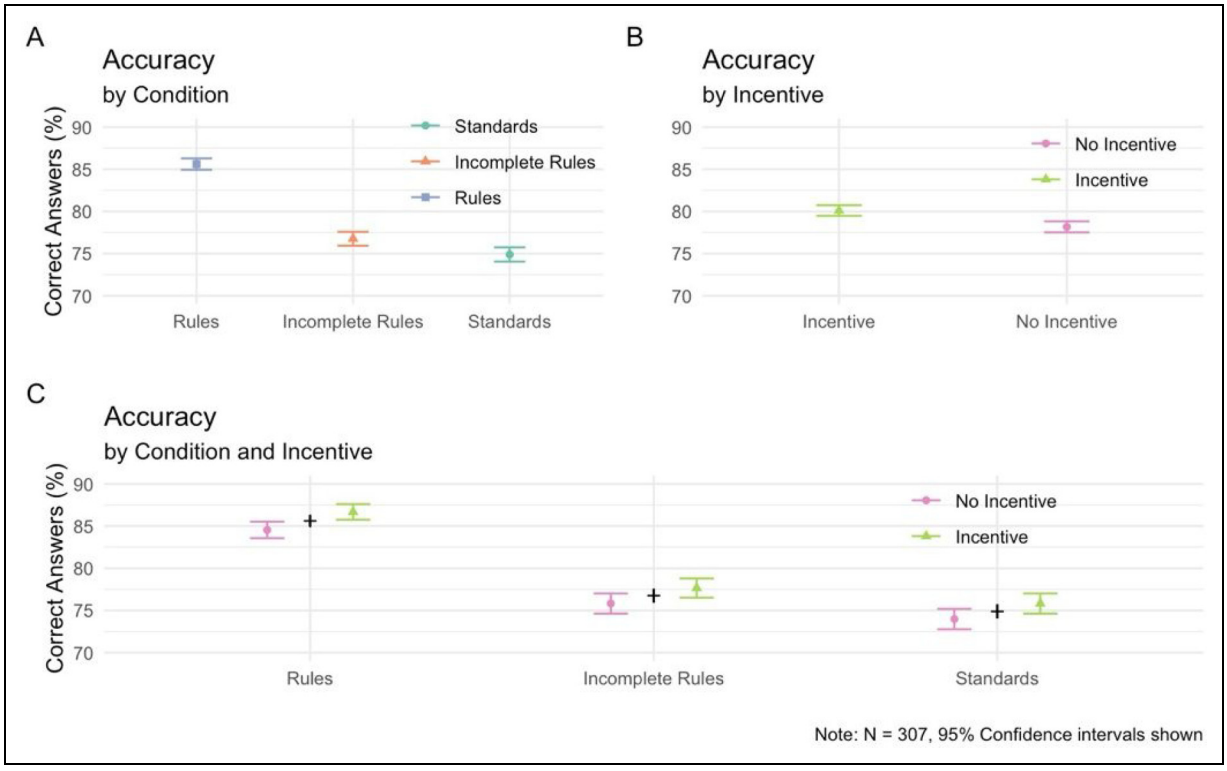

Figure 2. (A) Annotators working under rules have the highest accuracy rate (85.6%) of all participants, outperforming annotators working under incomplete rules (76.8%) and standards (74.9%). (B) Annotators working with the monetary incentive have a higher accuracy rate (80.1%) than those working without it (78.2%). (C) Annotators working under rules and with the monetary incentive had the highest accuracy rate of all participants (86.7%). Source: Laux et al. (2023). Note: The black crosses indicate the overall mean of each of the three conditions including their respective 95% confidence intervals.

images–especially when both conditions are combined, and annotators are working under rules *and* with a monetary incentive. However, the effects of a change in instructions on accuracy are larger than the effects of a change in pay. Although the effects of a monetary incentive on accuracy are marginal, the incentive still increased accuracy in the already high-performing rule conditions. In a competitive AI economy, AI developers may thus value even marginal improvements of annotations and, hence, marginally better training data for their models.

That said, the (industry-funded) Partnership on AI suggests paying data annotators 'at least the living wage for their location' (Partnership on AI, n.d.). Considering that much data annotation work is outsourced to regions with comparatively low salaries in the Global South, the additional costs of raising annotators' pay above the local living wage do not seem prohibitively high, especially for large tech companies (and may be offset by gains in accuracy). Additionally, organisations should consider how helpful their instructions are for annotators, and whether they can increase annotators' wellbeing by providing clearer guidance. The distinction between rules and standards provides an economic framework for testing guidance instructions.

While the results in Laux et al. (2023) have limited external validity–standards may, for example, yield better results in other annotation contexts–the insight that improving work conditions in data annotation positively impacts the quality of annotation outcomes and, thus, AI models holds implications for AI data work and human labour in the AI economy more generally. There is an emerging ethical and regulatory demand of maintaining human oversight of AI to mitigate risks to health, safety or fundamental rights (Sterz et al., 2024; Laux, 2024; Green, 2022). In as much as the process of annotation impacts the training of AI models–for example, regarding its ability to predict the location of curb ramps in cities for wheelchair users (Deitz, 2023)—it is relevant for mitigating AI's risks to fundamental rights. Publishing annotation instructions should thus be a precondition for scientific review (Rädsch et al., 2023; Gebru et al., 2018) and–in a regulatory context–independent audits and human oversight of AI systems. Moreover, under Article 14 of the European Union AI Act, AI developers will be obliged to instruct AI users on how to implement human oversight (Laux and Ruschemeier, 2025). AI policymakers should require testing on whether these instructions are best formulated as rules or standards.

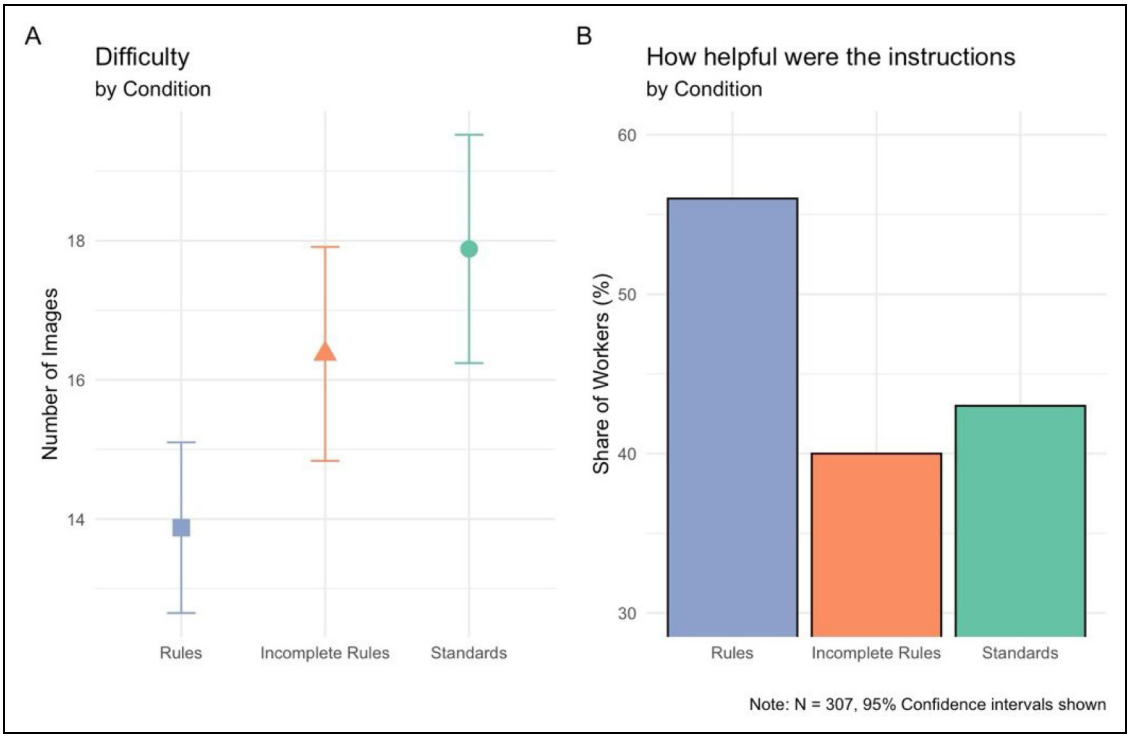

Figure 3. (A) Annotators working under rules on average report fewer images they had trouble with classifying (13.9) than those working under incomplete rules (16.4) and standards (17.9). (B) Annotators working under rules report their instructions more often to be very helpful (57%) than those working under incomplete rules (40%) and standards (44%). Source: Laux et al. (2023).

ORCID iDs

Johann Laux https://orcid.org/0000-0003-3043-075X
Fabian Stephany https://orcid.org/0000-0002-0713-6010
Alice Liefgreen https://orcid.org/0000-0001-8580-6924

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the British Academy, (grant number PF22 \220076).

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Braesemann F, Stephany F, Teutloff O, et al. (2022) The global polarisation of remote work. *PloS One* 17(10): e0274630.

Casey AJ and Niblett A (2017) The death of rules and standards. *Indiana Law Journal* 92(4): 1401–1447.

Chen Y, Clayton EW, Novak LL, et al. (2023) Human-centered design to address biases in artificial intelligence. *Journal of Medical Internet Research* 25: e43251.

Crawford K (2021) *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.

Deitz S (2023) Outlier bias: AI classification of curb ramps, outliers, and context. *Big Data & Society* 10(2): 20539517231203669.

Denton E, Hanna A, Amironesei R, et al. (2021) On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society* 8(2): 205395172110359.

Eisenmann M, Reinke A, Weru V, et al. (2023) *Why is the Winner the Best?*. https://doi.org/10.48550/ARXIV.2303.17719.

Gebru T, Morgenstern J, Vecchione B, et al. (2018) *Datasheets for Datasets*. https://doi.org/10.48550/ARXIV.1803.09010.

Gillier T, Chaffois C, Belkhouja M, et al. (2018) The effects of task instructions in crowdsourcing innovative ideas. *Technological Forecasting and Social Change* 134: 35–44.

Gonzalez Ehlinger E and Stephany F (2023) Skills or degree? The rise of skill-based hiring for AI and green jobs. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4603764

Graham M and Anwar MA (2019) The global gig economy: Towards a planetary labour market? *First Monday*. https://doi.org/10.5210/fm.v24i4.9913

Gray ML and Suri S (2019) *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass. Boston: Houghton Mifflin Harcourt.*

Green B (2022) The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review* 45: 105681.

Kaplow L (1992) Rules versus standards: An economic analysis. *Duke Law Journal* 42(3): 557–629.

Kaplow L and Shavell S (2002) Economic analysis of law. In: *Handbook of Public Economics*. Amsterdam: Elsevier, 3, 1661–1784. https://doi.org/10.1016/S1573-4420(02)80029-5

Korobkin RB (2000) Behavioral analysis and legal form: Rules vs. standards revisited. *Oregon Law Review* 79(1): 23–60.

Laux J (2024) Institutionalised distrust and human oversight of artificial intelligence: Towards a democratic design of AI governance under the European Union AI Act. *AI & SOCIETY* 39(6): 2853–2866. https://doi.org/10.1007/s00146-023-01777-z

Laux J and Ruschemeier H (2025) *Automation Bias in the AI Act: On the Legal Implications of Attempting to De-Bias Human Oversight of AI*. https://doi.org/10.2139/ssrn.5117560

Laux J, Stephany F and Liefgreen A (2023) *Improving Task Instructions for Data Annotators: How Clear Rules and Higher Pay Increase Performance in Data Annotation in the AI Economy* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2312.14565

Le Ludec C, Cornet M and Casilli AA (2023) The problem with annotation. Human labour and outsourcing between France and Madagascar. *Big Data & Society* 10(2): 20539517231188723.

Litman L, Robinson J and Rosenzweig C (2015) The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on mechanical Turk. *Behavior Research Methods* 47(2): 519–528.

Miceli M and Posada J (2022) *The Data-Production Dispositif* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2205.11963

Muldoon J, Cant C, Wu B, et al. (2024a) A typology of artificial intelligence data work. *Big Data & Society* 11(1): 20539517241232632.

Muldoon J, Graham M and Cant C (2024b) *Feeding the Machine*. Edinburgh: Canongate.

Northcutt CG, Athalye A and Mueller J (2021) *Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks*. https://doi.org/10.48550/ARXIV.2103.14749

Ørting SN, Doyle A, Van Hilten A, et al. (2020) A survey of crowdsourcing in medical image analysis. *Human Computation* 7: 1–26.

Partnership on AI (n.d.) *Improving Conditions for Data Enrichment Workers: Resources for AI Practitioners*. Retrieved 3 October 2023, from https://partnershiponai.org/responsible-sourcing-library/

Patra GK, Bhimala KR, Marndi A, et al. (2023) Deep learning methods for scientific and industrial research. In: *Handbook of Statistics*. London: Elsevier, 48, 107–168. https://doi.org/10.1016/bs.host.2022.12.002

Paullada A, Raji ID, Bender EM, et al. (2020) *Data and Its (Dis)contents: A Survey of Dataset Development and Use in Machine Learning Research*. https://doi.org/10.48550/ARXIV.2012.05345

Rädsch T, Reinke A, Weru V, et al. (2023) Labelling instructions matter in biomedical image analysis. *Nature Machine Intelligence* 5(3): 273–283.

Shemtob L, Beaney T, Norton J, et al. (2023) How can we improve the quality of data collected in general practice? *BMJ* 380: e071950. https://doi.org/10.1136/bmj-2022-071950

Stephany F and Teutloff O (2024) What is the price of a skill? The value of complementarity. *Research Policy* 53(1): 104898.

Sterz S, Baum K, Biewer S, et al. (2024) *On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives*. https://doi.org/10.48550/ARXIV.2404.04059

Tubaro P (2021) Disembedded or deeply embedded? A multi-level network analysis of online labour platforms. *Sociology* 55(5): 927–944.

Tubaro P, Casilli AA and Coville M (2020) The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence. *Big Data & Society* 7(1): 205395172091977.

Wang H, Fu T, Du Y, et al. (2023) Scientific discovery in the age of artificial intelligence. *Nature* 620(7972): 47–60.

Wong M (2023) America already has an AI underclass. *The Atlantic*. https://www.theatlantic.com/technology/archive/2023/07/ai-chatbot-human-evaluator-feedback/674805/

Wood AJ, Graham M, Lehdonvirta V, et al. (2019) Good gig, bad gig: Autonomy and algorithmic control in the global gig economy. *Work, Employment and Society* 33(1): 56–75.

Yang K, Qinami K, Fei-Fei L, et al. (2020) Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 547–558. https://doi.org/10.1145/3351095.3375709