Vol. 3, No. 4, pp.: 1-4, 2025

The Data Labelling Crisis: Why Ground Truth Is Becoming AI's Scarcest Resource

Amira Khalil[1,*], **Thomas van der Berg**[2]

[1] Oxford Internet Institute, University of Oxford, Oxford, UK, OX1 3JS

[2] Department of Information and Computing Sciences, Utrecht University, Utrecht, Netherlands, 3584 CS

[*] *amira.khalil@oii.ox.ac.uk*

Article Information	
Received	2 May 2025
Accepted	24 December 2025
DOI	https://doi.org/10.63646/datamind.2025.030401

Abstract

The AI research community's relationship with labelled data has always been ambivalent — it is simultaneously the foundation of supervised learning and the bottleneck that constrains what supervised learning can achieve. As large-scale pre-training has reduced the labelled data requirements for many standard NLP and computer vision tasks, a somewhat premature narrative has emerged suggesting that the labelling problem is essentially solved. This perspective argues the opposite. The labelling crisis is not receding; it is intensifying and shifting. The tasks for which labelling is most needed — nuanced safety evaluation, cross-cultural preference alignment, specialised scientific annotation, temporal event ordering in long documents — are precisely those for which automated labelling approaches perform worst, crowd-sourcing quality is most unreliable, and the epistemic foundations of 'ground truth' are most contested. We examine three dimensions of the labelling crisis: the labour economics of annotation at scale, the epistemic problem of ground truth in subjective or culturally-dependent tasks, and the emerging challenge of synthetic data for LLM training as a substitute for — and complication of — human-generated ground truth. We conclude with specific recommendations for the research community and for organisations building annotation infrastructure.

Keywords: ***data annotation; ground truth; labelling; crowdsourcing; RLHF; synthetic data; data workers; AI safety***

1. A Crisis in Plain Sight

The ImageNet database, when it launched in 2009, contained 14 million images labelled by approximately 50,000 Amazon Mechanical Turk workers over several years. It took three years and a reported investment of around $500,000 to produce a dataset that transformed computer vision. Today, the most capable multimodal AI systems are trained on datasets orders of magnitude larger, with labelling tasks that are substantially more complex — safety evaluations, preference rankings, multi-step reasoning annotations — and with quality requirements that are significantly higher. The economics are simply not scaling.

The optimistic view is that self-supervised and semi-supervised learning have reduced the dependency on labelled data dramatically, and that synthetic data generated by models themselves can fill the gap. This view is partially correct for some tasks and wildly incorrect for others. A sentiment classifier needs far fewer labelled examples than it did in 2015, thanks to pre-trained

representations. A model for evaluating whether an AI assistant's response to a sensitive mental health query is safe, compassionate, and culturally appropriate — that still needs human annotators, and a lot of them, and not just any human annotators.

2. The Labour Economics of Annotation

The annotation economy has been well-documented by researchers like Irani and Silberman, who coined the term 'labour-centric design' in the context of crowdwork, and more recently by TIME's investigation into the data labellers behind ChatGPT's safety training in Kenya. The structural features of this economy are troubling. Annotation work is typically performed by workers in lower-income countries, paid at rates of $1–5 per hour, with minimal job security and limited recourse for poor working conditions. The psychological cost of safety annotation work — which requires repeated exposure to harmful, violent, or disturbing content — is substantial and largely unaddressed by the companies that commission it.

This is not just an ethical problem (though it is certainly that). It is also a quality problem. Workers performing repetitive, low-paid annotation tasks under time pressure are not producing the nuanced, consistent labels that model safety and alignment work requires. The incentive structures of crowdwork platforms create a systematic pressure toward speed over accuracy that is poorly controlled by the quality assurance mechanisms typically deployed. The result is labelled data of variable and often unknowable quality, used as ground truth for systems making consequential decisions.

3. The Epistemic Problem of Ground Truth

The phrase 'ground truth' implies an objective correctness that is conceptually appropriate for some annotation tasks (is there a cat in this image?) and deeply inappropriate for others (is this response to a user's political question balanced? Is this content inappropriate for children?). The epistemic status of the label is not a secondary concern; it determines what the model is actually learning and how its performance should be evaluated.

The field's standard treatment of annotation disagreement — compute inter-annotator agreement, discard low-IAA examples, take the majority vote — papers over a fundamental conceptual problem. For culturally sensitive tasks, annotation disagreement is often not noise to be filtered but signal to be preserved. A content moderation classifier trained on American annotator labels will encode American cultural norms about acceptable speech. Majority-vote aggregation does not produce a universally correct label; it produces the majority's label. Calling it 'ground truth' and training a global system on it has consequences that the current annotation paradigm does not adequately address.

[Figure 1 — Spectrum of annotation tasks by epistemic status of ground truth (horizontal axis) and scalability of automated labelling approaches (vertical axis). Tasks in the lower-right quadrant (high objectivity, high automation) include standard image classification and named entity recognition. Tasks in the upper-left quadrant (low objectivity, low automation) — nuanced safety evaluation, cultural preference alignment, cross-domain scientific annotation — represent the tasks where the labelling crisis is most acute. These are also the tasks with the most immediate stakes for AI safety and alignment.]

Figure 1. *Spectrum of annotation tasks by epistemic status of ground truth (horizontal axis) and scalability of automated labelling approaches (vertical axis). Tasks in the lower-right quadrant (high objectivity, high automation) include standard image classification and named entity recognition. Tasks in the upper-left quadrant (low objectivity, low automation) — nuanced safety evaluation, cultural preference alignment, cross-domain scientific annotation — represent the tasks where the labelling crisis is most acute. These are*

also the tasks with the most immediate stakes for AI safety and alignment.

4. Synthetic Data as Supplement and Complication

The use of LLM-generated synthetic data to supplement or replace human annotation is simultaneously the most promising near-term response to the labelling bottleneck and a source of new epistemic complications. For tasks where an LLM can generate high-quality synthetic examples that are indistinguishable from human-labelled examples — paraphrase generation, data augmentation for low-resource languages, generating adversarial test cases — synthetic data is a genuine productivity multiplier. For tasks where label quality is the problem — safety evaluation, cultural sensitivity, preference alignment — using an LLM to generate labels risks systematically encoding the LLM's existing biases into the training labels for the next generation of models.

This is not a hypothetical risk. RLHF-trained models are already widely used to generate preference labels for LLM alignment training. If those models have systematic biases in their safety or preference judgments, those biases will be amplified in the models trained on their outputs. The 'model collapse' literature has documented one dimension of this problem — training on synthetic data degrades distributional diversity. The alignment version of this problem, where training on synthetic preference labels degrades alignment diversity, is less well-studied and potentially more consequential.

5. Recommendations

We offer three recommendations. First, invest in annotation infrastructure that treats annotator diversity as a design requirement, not an afterthought. For tasks where cultural and demographic perspective matters, intentional diversity in annotator populations — not incidental diversity through low-wage crowdwork — is necessary. Second, develop and adopt multi-label annotation formats that preserve disagreement rather than resolving it to majority votes, and build evaluation frameworks that assess model performance across the distribution of annotator views. Third, treat synthetic data annotation the same way the field treats automated evaluation: as a tool that is appropriate for some tasks and problematic for others, and that requires empirical validation against human labels rather than assumption of equivalence.

6. Conclusion

Ground truth is not given. It is produced — by specific people in specific contexts with specific incentives and constraints. The AI field's tendency to treat annotation as a solved, commoditised, infrastructure problem has produced data quality issues that are baked into some of the most widely-deployed AI systems. Addressing the labelling crisis requires engaging with its labour economics, its epistemic foundations, and its interaction with synthetic data in ways that the current research paradigm largely avoids. The conversation is overdue.

References

Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021). On the genealogy of machine learning datasets: A critical history of ImageNet. Big Data & Society, 8(2). https://doi.org/10.1177/20539517211035955

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis)contents: A survey

of dataset development and use in machine learning research. Patterns, 2(11), 100336. https://doi.org/10.1016/j.patter.2021.100336

Hare, B. (2018). Crowdwork, invisible and on demand. International Journal of Communication, 12, 3934–3954. https://ijoc.org/index.php/ijoc/article/view/6401

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744. https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2024). The curse of recursion: Training on generated data makes models forget. arXiv preprint arXiv:2305.17493. https://doi.org/10.48550/arXiv.2305.17493

Aroyo, L., & Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. AI Magazine, 36(1), 15–24. https://doi.org/10.1609/aimag.v36i1.2564

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? FAccT 2021, 610–623. https://doi.org/10.1145/3442188.3445922

Northcutt, C. G., Athalye, A., & Mueller, J. (2021). Pervasive label errors in test sets destabilize machine learning benchmarks. NeurIPS 2021 Datasets Track. https://proceedings.neurips.cc/paper/2021/hash/f2217d0ec136b5a4b60580c1e6bfc027-Abstract-Datasets_and_Benchmarks.html

Gray, M. L., & Suri, S. (2019). Ghost work: How to stop Silicon Valley from building a new global underclass. Houghton Mifflin Harcourt. https://ghostwork.info/

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92. https://doi.org/10.1145/3458723