Annotation Practices in Societally Impactful Machine Learning Applications

What are these automated systems actually trained on?

– A NeurIPS Case Study –

Introduction

Machine Learning (ML) models rely on labelled datasets for training and evaluation. These labels, the "ground truth", directly affect model accuracy and trustworthiness. However, ML research emphasizes performance and novelty [1], while data and labelling remain under-reported [2]. Poor annotation practices and transparency risk biased, unreliable, or unreproducible results.

As ML becomes more embedded in society, dataset creation and labelling must face greater scrutiny to ensure the integrity of its applications.

NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS is one of the leading ML conferences, by h5-index [3]. It covers a wide range of topics, including deep learning, reinforcement learning, and fairness in ML. In 2020, a section addressing broader societal impact was required in NeurIPS papers. Since 2021, it was no longer mandatory, but remained encouraged through author instructions. As such, it is worth assessing how ground truths are approached in discussions about societal impact, if at all.

Research Question

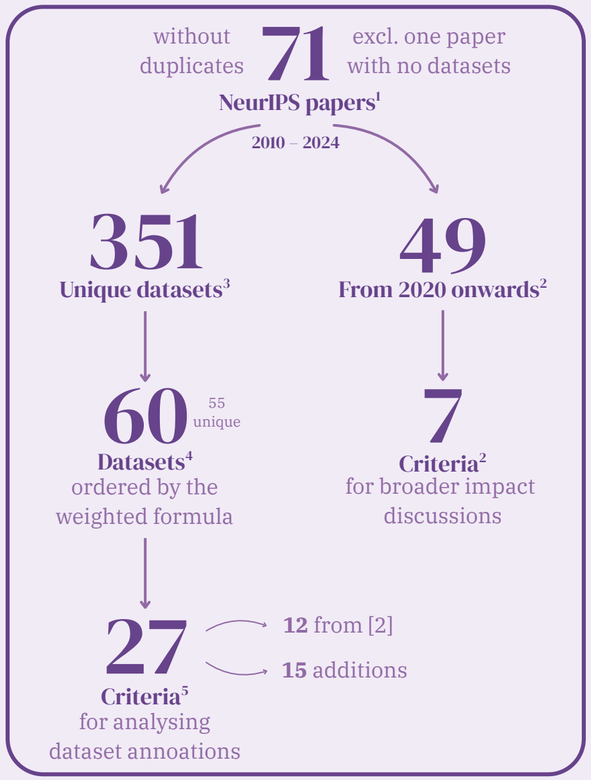
"What are the data collection and reporting practices for annotation in societally impactful ML applications from the NeurIPS venue?"

- **SQ1** How do NeurIPS researchers assess the quality of the datasets that they use for their models? Do they explicitly take annotations into account?
- **SQ2** What or who is labelling the datasets?
- **SQ3** What are the relevant criteria for evaluating the transparency of dataset creation?
- **SQ4** Do the datasets fit the criteria established by SQ3?

Methodology

1. The 25 highest cited NeurIPS papers from each of the last 2, 5, and 15 years are queried from Scopus.
2. "Broader Impact" sections of the papers from 2020 onwards are inspected to observe how annotations are perceived in this context.
3. Datasets employed in the 75 papers are extracted.
4. A weighted citaion-based formula* is used to get the top 20 datasets from each time frame.
5. These are assessed using criteria from past literature [2] and new additional indicators.

*citation sum of papers from that period that use the dataset



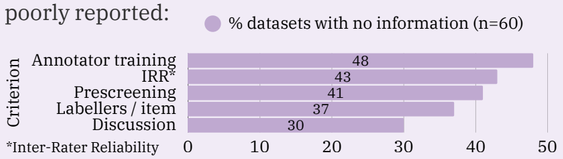
Main Results

n=60

91% of datasets had a reliable source of information about them.

65% were either fully or partially labelled by humans.

However, procedural details regarding annotation are poorly reported:



Authors generally prefer formal instructions instead of interactive training & discussions. Yet this has the potential to worsen ground truths by making them devoid of nuance [4]. Moreover, omitting IRR puts the reliability of the labels under doubt. Temporal improvements wrt. documentation are not substantial.

89% of datasets describe their item population, but there are cases of web-crawled visual data that is not inspected. Independent audits have discovered unethical contents in such datasets [5].

n=49

69% of NeurIPS papers (2020 onwards) contained a section discussing societal impact.

of which **68%** do **not** address annotation quality.

NeurIPS researchers often do not link the impact of ML systems on humans and the fact that ground truths are a fundamental component of said systems.

35% of studies use datasets for model training / evaluation that are **not** made public.

This undermines trust in the proposed models, as third parties cannot assess data or label quality. It is especially concerning as a portion of the 35% of papers were authored by strong ML industry players, whose products reach millions. Moreover, it raises doubts about the transparency standards of the venue.

Conclusions

I. While high-level information about datasets is available, the annotation process remains poorly reported.

II. Large-scale visual datasets are uncurated and prone to containing harmful data.

III. NeurIPS research refrains from relating ground truth quality to model quality and positive impact.

IV. There is a concerning amount of datasets that are not publicly available, with no information about annotations, which implies questionable reproducibility norms.

Future Work

- Call for **standardisation** of reporting practices on annotation, as part of transparency norms.
- NeurIPS needs to ensure that datasets are a **publicly accessible asset** as part of the same standads.
- More **collaborative efforts** between authors and annotators, to ensure reliability of labels and data. **OpenAssistant Conversations** is a project that shows how this can be achieved even with limited funding, and it comes from within NeurIPS [6].
- Dataset content should face more scrutiny from the ML research community as an effort to **minimise potential harm**.

Contact Information

Author: Simona Cristina Lupșa, TU Delft, 2025.
s.c.lupsa@student.tudelft.nl

Supervisors: Dr. Cynthia Liem, Andrew M. Demetriou

References: **[1]** A. Birhane, P. Kalluri, D. Card, W. Agnew, R. Dotan, and M. Bao, "The values encoded in machine learning research," in Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, (New York, NY, USA), pp. 173–184, Association for Computing Machinery, 2022. **[2]** R. S. Geiger, D. Cope, J. Ip, M. Lotosh, A. Shah, J. Weng, and R. Tang, " "garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data?," Quantitative Science Studies, vol. 2, pp. 795–827, 11 2021. **[3]** Google, "Top publications – google scholar metrics." https://scholar.google.com/citations?view_op=top_venues, 2025. **[4]** L. Aroyo and C. Welty, "Truth is a lie: Crowd truth and the seven myths of human annotation," AI Mag., vol. 36, pp. 15–24, Mar. 2015. **[5]** A. Birhane and V. Prabhu, "Large image datasets: A pyrrhic win for computer vision?," pp. 1536–1546, 01 2021. **[6]** A. Köpf, Y. Kilcher, D. von Rütte, S. Anagnostidis, Z. R. Tam, K. Stevens, A. Barhoum, D. Nguyen, O. Stanley, R. Nagyfi, S. ES, S. Suri, D. Glushkov, A. Dantuluri, A. Maguire, C. Schuhmann, H. Nguyen, and A. Mattick, "Openassistant conversations - democratizing large language model alignment," in Advances in Neural Information Processing Systems, vol. 36, pp. 47669–47681, 2023.