Values Encoded in Machine Learning Research

Isabel Kurth*

January 29, 2025

Abstract

1 Introduction

Machine learning (ML) research has become a pivotal driver of technological advancement, shaping fields as diverse as healthcare, finance, and autonomous systems. Despite its transformative potential, the values embedded in influential ML research remain underexamined, raising critical questions about whom the field serves and whose needs it prioritizes. Birhane et al. (2023) [1] underscore that ML research often prioritizes performance, generalization, and novelty while neglecting societal needs, ethical considerations, and broader societal impact.

Inspired by this work, our study systematically investigates the values reflected in the 100 most influential papers from NeurIPS 2023 and ICML 2023. These conferences represent the cutting edge of ML research and are key arenas where emergent trends and priorities are crystallized. By analyzing these highly-cited papers, we aim to uncover patterns in their value commitments, especially regarding how they balance technical rigor with societal impact. This work aims to investigate if the most encoded values changed from the years 2008/09 and 2018/19 (investigated by Birhane et al. (2023)) to more recent papers from 2023.

This paper contributes to the growing discourse on the socio-political dimensions of ML research by providing empirical insights into the values upheld in contemporary work. Through our analysis, we aim to encourage researchers, institutions, and policymakers to critically reflect on the long-term implications of prioritizing certain values over others.

*Hasso Plattner Institute, Potsdam, Germany. isabel.kurth@student.hpi.de

2 Methodology

2.1 Data collection

We curated a dataset of the 100 most influential papers from NeurIPS 2023 and ICML 2023 using rankings provided by Paper Digest (`https://www.paperdigest.org/topic/?topic=nips&year=2023` and `https://www.paperdigest.org/topic/?topic=icml&year=2023`). These rankings were based on an impact score, which reflects a combination of paper citations, patent citations, etc. It is a score in 1.0-10.0, with a higher value indicating a broader impact. We then downloaded the respective papers from the conference website directly (`https://proceedings.neurips.cc/paper_files/paper/2023` and `https://icml.cc/Downloads/2023`). The selected papers represent the cutting edge of machine learning research and provide a representative sample of current trends and values in the field. We did not use Semantic Scholar as in the Birhane et al. (2023) paper because Semantic Scholar did not provide API keys due to high demand. We could also not use Google Scholar to access the citation counts because the IP address gets blocked after too many access tries. A list of the PDFs of the 100 papers of both conferences can be found in the GitHub repository.

2.2 Keyword based analysis

To analyze the values encoded in these papers, we adopted the 74 keywords used by Birhane et al. (2023) in their study of influential ML papers. In their paper they write that they use 67 values, but in their annotation template for manual annotations (annotations.tsv) they include 74 values. In their results folder they just mention 62 values. In order to search for as many values as possible we decided to proceed with the 74 values from the annotation file in their data folder. These keywords include terms related to performance, generalization, novelty, societal impact, and ethical considerations. Unlike Birhane et al., who relied on manual annotation, we developed and deployed an automated keyword-scanning algorithm to systematically identify the presence of these keywords across the abstracts, introductions, and conclusions of the selected papers. The Table1 in the Appendix provides a comprehensive list of the values and their associated keywords used in our analysis. For example we did not only query for the additonal keyword “Novelty” but also for “novel” as these two words express the same value and because we use a keyword search, sentences with the word “novel” would not be counted towards the keyword “novelty”. For each keyword occurrence, contextual information was extracted to ensure accurate interpretation. This semi-automated process balanced the scale of automated analysis with the depth of manual review.

2.3 Quantitative Summary

We measured the prevalence of each keyword in the dataset, calculating the relative frequency of each value.

2.4 Qualitative Summary

Excerpts containing the keywords were manually reviewed to explore how values are framed and justified. Particular attention was given to the framing of societal impacts and ethical considerations, drawing comparisons with prior findings by Birhane et al.

References

[1] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. The values encoded in machine learning research. *arXiv preprint arXiv:*, 2021.

A Keyword Analysis Table

Keywords Birhane Paper	additional keywords for automated keyword search
Novelty	novel
Simplicity	simple
Generalization	generalisable
Flexibility/Extensibility	flexible, flexibility, extensibility, extensible
Robustness	robust
Realistic output	realistic
Formal description/analysis	formal, mathematical
Theoretical guarantees	guarantee
Approximation	approximate
Quantitative evidence (e.g. experiments)	experiment
Qualitative evidence (e.g. examples)	example
Scientific methodology	scientific
Controllability (of model owner)	control
Human-like mechanism	human
Low cost	cheap, cost
Large scale	scale
Generality	general
Principled	principles
Exactness	exact
Preciseness	precise
Concreteness	correct
Automatic	automated
Efficiency	efficient
Building on classic work	classic work
Building in recent work	recent work
Unifying ideas or integrating components	unifying
Identifying limitations	limitations
Critique	criticism
Understanding (for researchers)	understanding
Used in practice/Popular	practice, popular
Reproducibility	reproduce
Easy to implement	implement
Requires few resources	resources
Parallelizability / distributed	parallelizability, parallelization, distributed
Facilitating use (e.g. sharing code)	sharing code
Scales up	scale up
Applies to real world	real world
Learning from humans	humans, learning
Practical	practice
Useful	usefulness
Interpretable (to users)	interpretable
Transparent (to users)	transparent, transparency
Privacy	privacy, private
Fairness	fair
Not socially biased	social bias, socially bias, social, society
User influence	user
Collective influence	collective
Deferral to humans	deferral
Critiqability	criticism
Beneficence	beneficable
Respect for Persons	respect
Autonomy (power to decide)	autonomy, autonome
Explicability	explicable
Respect for Law and public interest	respect for law, respect for public interest
Security	secure