Examining Axiological Assumptions in Machine Learning Publications

Yashpreet Malhotra
yashmalhotra9323@gmail.com

***Abstract*—This paper presents a study of the values embedded within machine learning research papers. A novel annotation scheme is developed to analyze how values are represented in scholarly documents, focusing on the rationales for research projects, the emphasized attributes of those projects, and the discussion or neglect of potential negative impacts. The methodology is applied to a corpus of influential papers from top-tier machine learning conferences. The analysis explores the relationship between these encoded values and factors such as institutional affiliations and funding sources, aiming to contribute to a more nuanced understanding of the ethical dimensions of machine learning research.**

***Index Terms*—Machine learning research, scientific values, ethical analysis, research rationales, negative impacts, institutional affiliations, funding sources, scholarly discourse.**

I. Introduction

Over the past decade, machine learning (ML) has evolved from a niche academic discipline into a cornerstone of technological advancement, influencing a wide array of industries including healthcare, finance, transportation, and social media. Its rapid proliferation has been accompanied by increasing public attention, government investment, and corporate interest. As ML systems become more deeply integrated into the fabric of society, concerns surrounding their ethical, social, and political implications have gained prominence. However, despite these growing concerns, mainstream ML research often continues to treat such issues as peripheral rather than central to the development process.

A dominant narrative persists within the research community that presents ML development as an inherently technical and neutral endeavor—driven by objectivity, performance benchmarks, and empirical rigor. This framing, while useful for fostering scientific progress, tends to obscure the fact that every design choice, evaluation criterion, and research focus reflects an underlying set of values. The choices researchers make—what problems to tackle, which metrics to optimize, whose data to use, and what outcomes to prioritize—are never fully divorced from societal context. Yet, these considerations are frequently underexplored in published literature, which often defaults to a narrow set of values such as accuracy, scalability, and novelty.

This study aims to interrogate the implicit value systems embedded in the corpus of ML research. Specifically, it seeks to uncover which attributes are consistently praised, which concerns are systematically excluded, and how these patterns reflect broader institutional forces shaping the field. By conducting a qualitative and thematic analysis of 100 highly cited papers published at top-tier venues—specifically the International Conference on Machine Learning (ICML) and the Conference on Neural Information Processing Systems (NeurIPS)—between 2008 and 2019, this work offers a critical lens into the internal logic of the ML research ecosystem.

These papers were selected not only for their influence, as measured by citation count, but also for their role in setting research agendas and defining norms within the field. The analysis investigates the extent to which the values promoted in these works align with broader societal goals, and how institutional affiliations—whether academic, corporate, or hybrid—affect the framing of research contributions. In doing so, this study contributes to ongoing conversations about the social responsibility of ML research and calls for a more reflective and inclusive approach to innovation in the field.

II. Methodology

To systematically investigate the implicit and explicit value structures embedded within machine learning research, we adopted a multi-phase interpretive methodology grounded in both qualitative analysis and rigorous validation protocols. This hybrid approach enabled a nuanced examination of how research in ML is framed, what kinds of justifications are offered for methodological choices, and which societal considerations are emphasized, downplayed, or omitted altogether.

A. Corpus Selection

The dataset consisted of 100 high-impact research papers selected from two of the most prestigious venues in the field of machine learning: the International Conference on Machine Learning (ICML) and the Conference on Neural Information Processing Systems (NeurIPS). These conferences were chosen for their longstanding influence and for setting research trends within the global ML community. Papers were drawn from four strategically chosen years—2008, 2009, 2018, and 2019—thereby capturing both the early development of modern ML techniques and their more recent, industrially-integrated manifestations. Citation count was used as a proxy for impact, ensuring that the papers analyzed were widely read and influential in shaping discourse.

B. Annotation Framework and Process

The textual content of each paper was segmented at the sentence level across four core sections: Abstract, Introduction, Discussion, and Conclusion. These sections were chosen because they encapsulate the narrative arc of the paper—from

motivation and framing to the articulation of contributions and broader implications.

A team of trained annotators conducted a detailed manual coding of each sentence using a hybrid coding framework that combined deductive and inductive strategies. The deductive component was informed by existing literature on responsible AI and technology ethics, focusing on established normative categories such as fairness, transparency, accountability, safety, and social benefit. This provided a structured lens through which to assess the explicit values referenced in each paper.

In parallel, the inductive approach allowed for the emergence of novel themes not predefined in the coding schema. This was essential for capturing domain-specific justifications or subtle rhetorical strategies that may reflect implicit values. Annotators were instructed to identify and classify sentence-level value statements according to several dimensions: the type of value elevated (e.g., novelty, performance, efficiency), the presence or absence of societal impact acknowledgment, and the rhetorical justification strategies employed (e.g., appeals to utility, objectivity, scalability).

C. Validation and Reliability Measures

To ensure analytical rigor and mitigate subjectivity, 40% of the dataset underwent dual annotation, followed by a reconciliation process to address discrepancies. The inter-annotator agreement was assessed using both raw percentage agreement and statistical metrics. An overall agreement rate of 87% was achieved, demonstrating a high level of consistency in the identification and categorization of value-laden statements.

In addition, a fuzzy Fleiss' kappa score of 0.45 was computed, suggesting moderate agreement across multiple annotators for categorical variables. For ordinal variables such as the degree of societal justification and acknowledgment of potential harms, a weighted Fleiss' kappa exceeding 0.6 was recorded, indicating substantial inter-rater reliability. These validation metrics reinforce the reliability of the annotation framework and support the credibility of subsequent analyses.

D. Institutional Attribution Analysis

Beyond textual analysis, each paper was reviewed for metadata relating to institutional affiliation and funding disclosures. Authors' primary affiliations were classified as academic, corporate, or hybrid (e.g., university-corporate collaborations). Publicly disclosed funding sources were also documented, with particular attention paid to corporate sponsorship or government grants. This layer of analysis was designed to examine whether institutional context influences the values promoted in ML research—either directly through funding priorities or indirectly through organizational norms and incentives.

By triangulating content-level findings with metadata on institutional context, this methodology offers a comprehensive lens through which to understand the value orientations embedded in mainstream ML research.

III. Quantitative Results

This section presents the quantitative findings derived from the analysis of 100 highly cited ML papers. The results are organized around key axes of ethical and societal engagement: justification of research relevance, acknowledgment of potential harms, value prioritization, and institutional trends.

A. Justification Distribution

To assess the extent to which authors justify the broader relevance of their research, each paper was classified into one of four categories based on the nature and depth of its societal justification.

TABLE I: Classification of Justification Strategies in Analyzed Papers

Justificatory Classification	Proportion of Papers
No Mention of Societal Relevance	68%
Stated but Unjustified Societal Link	17%
Minimal Societal Justification	11%
Detailed Societal Rationale	4%

As shown in Table I, a striking 68% of the analyzed papers made no reference to the societal implications or applications of their work. An additional 17% included vague or symbolic gestures toward societal relevance (e.g., claims of "real-world applicability") without substantiating such claims with detailed argumentation or evidence. Only 4% of papers offered a robust societal rationale—highlighting a significant gap in ethical contextualization across the literature.

B. Negative Implication Discussion

A separate coding exercise investigated the degree to which authors engaged with potential negative outcomes or ethical risks associated with their contributions.

TABLE II: Extent of Negative Impact Consideration in Sampled Literature

Engagement with Potential Harm	Proportion of Papers
No Recognition of Negative Outcomes	98%
Brief Mention of Possible Harms	1%
Substantive Risk Discussion	1%
In-Depth Harm Analysis	0%

Table II reveals that nearly all papers (98%) omitted any discussion of potential negative consequences arising from the proposed methodologies. Only 2 papers mentioned possible harms, and even these were restricted to cursory or high-level remarks. Not a single paper in the sample engaged in a deep or systematic exploration of unintended impacts—highlighting a profound asymmetry between technical enthusiasm and ethical foresight.

C. Figures and Interpretations

Figure 1 illustrates the dominant values emphasized across the dataset. Performance metrics such as accuracy, precision, and F1 score were by far the most cited benchmarks of success, followed closely by generalization ability and computational efficiency. Conversely, socially-relevant values such as fairness, interpretability, and safety were mentioned

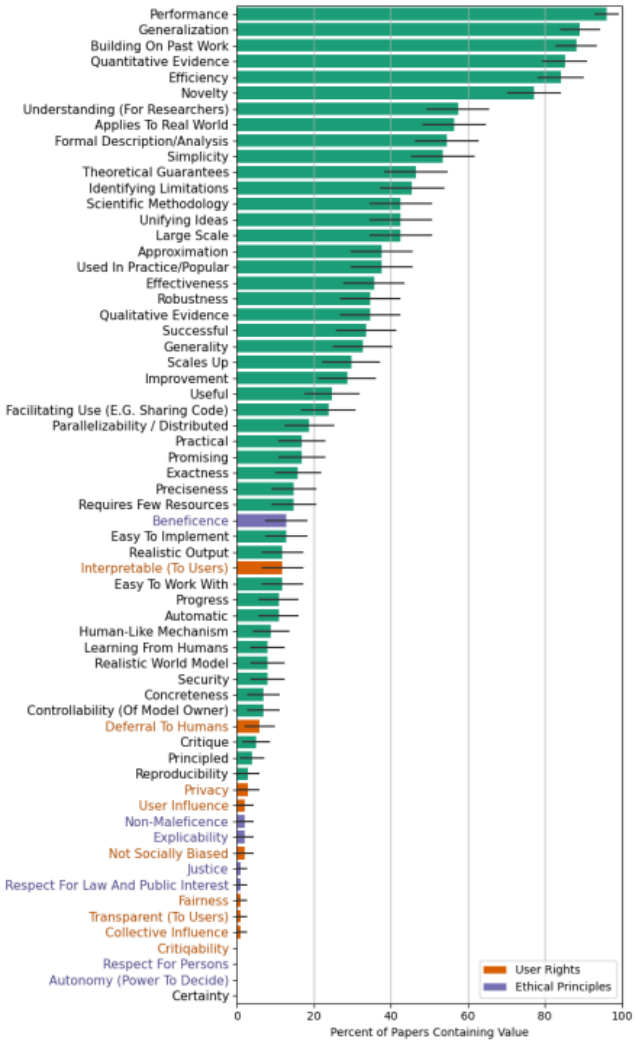

Fig. 1: Frequency distribution of top values (e.g., performance, generalization, efficiency) across reviewed papers.

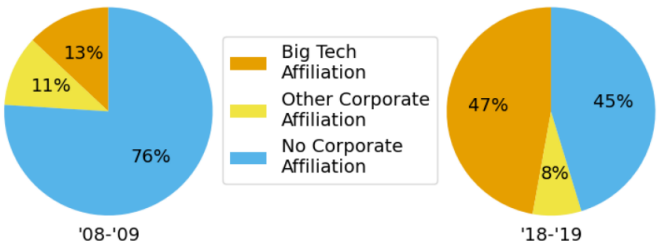

Fig. 2: Temporal shift in paper affiliations: Increased prevalence of Big Tech and elite academic institutions over time.

infrequently, indicating a skew toward optimization-centric paradigms.

Figure 2 highlights a notable temporal shift in the institutional makeup of ML authorship. Papers published in 2008 and 2009 were predominantly authored by researchers from academic institutions. However, by 2018 and 2019, the presence of large technology companies—particularly U.S.-based firms such as Google, Facebook, and Microsoft—had surged. This trend reflects the growing influence of corporate actors in setting research agendas, potentially reinforcing a value system oriented toward scalability, commercial applicability, and proprietary advantage.

IV. Value Analysis

A thematic exploration of the annotated corpus revealed six dominant values recurrently emphasized in the reviewed ML research: *Performance*, *Generalization*, *Efficiency*, *Building on Prior Work*, *Quantitative Evidence*, and *Novelty*. These were not only frequently mentioned but often operationalized in ways that prioritized technical benchmarks over social impact or inclusivity.

A. Performance

Performance was the most cited metric of value, often expressed through improvements in accuracy or superiority over existing methods. This emphasis generally revolved around benchmark datasets and quantitative evaluation metrics such as accuracy, F1 score, or mean squared error. Yet, such evaluations often lacked reflection on the limitations of the metrics used, particularly their inability to capture broader ethical or contextual concerns. Most papers treated accuracy as a proxy for progress, without discussing what such ”improvement” means for real-world stakeholders.

B. Generalization

Generalization was primarily framed as the capability of a model to perform consistently across unseen data or varying tasks. However, this concept was almost always evaluated using multiple curated datasets rather than deployment scenarios. Rarely did authors reflect on the sociotechnical implications of transferring models across domains, such as the potential amplification of biases or the ethical concerns of applying ML systems without contextual calibration.

C. Efficiency

Efficiency was commonly understood as reduced computational cost, memory usage, or training time. Paradoxically, many papers used the term ”efficient” to indicate the ability to scale to massive data or models—essentially enabling high-resource operations rather than reducing resource consumption. Energy use, environmental impact, or the democratization of access were virtually absent from the discussion, implying that the term “efficiency” primarily served large institutions with abundant resources.

D. Building on Prior Work and Novelty

Most papers sought to balance novelty with continuity by demonstrating how their contributions extended or refined existing methods. Novelty was typically framed in algorithmic or architectural terms, while social innovation or problem recontextualization was nearly invisible. Even when prior limitations were mentioned, they were addressed strictly from a performance lens, not through broader critical analysis.

V. Institutional and Corporate Influence

A detailed examination of author affiliations and funding disclosures across the sampled corpus reveals a pronounced concentration of influence among elite academic institutions and major technology corporations. These institutional forces play a pivotal role in shaping not only the direction of research but also the values and priorities embedded within the ML literature. Over the span from 2008 to 2019, this influence intensified, signaling broader structural shifts in the machine learning research ecosystem.

A. Authorship Trends

Analysis of authorship patterns across the selected papers demonstrates a marked increase in contributions from corporate-affiliated researchers, particularly those associated with multinational technology companies such as Google, Microsoft, Facebook, and Amazon. Between 2008–2009 and 2018–2019, the proportion of papers with at least one author affiliated with a corporate entity nearly tripled. This rise coincided with a growing dominance of elite academic institutions, including but not limited to Stanford, MIT, Carnegie Mellon, and UC Berkeley.

In contrast, the representation of authors from smaller universities, non-Western institutions, and under-resourced academic settings remained minimal to negligible. This stratification reflects a consolidation of research capital within a narrow band of institutions capable of providing substantial resources, infrastructure, and visibility—factors that significantly increase the likelihood of acceptance at top-tier conferences and high citation impact.

The increasing prevalence of industry-academic collaborations also merits attention. While such partnerships can yield powerful synergies, they often risk aligning academic inquiry with commercial imperatives, particularly when corporate entities contribute funding, data, or compute resources that are inaccessible to independent researchers.

B. Funding Patterns

A complementary analysis of funding acknowledgments (where disclosed) reveals a similarly skewed landscape. A substantial proportion of the papers did not explicitly state their sources of financial support. However, through cross-referencing institutional affiliations and known partnerships, we inferred that corporate backing—either directly or indirectly—played a significant role in supporting many of the most-cited works.

Among papers that did disclose funding information, industry sponsors were disproportionately represented. Funding from companies involved in the development and deployment of machine learning systems—especially those with commercial interests in scalability, speed, and competitive advantage—was widespread. This trend suggests a potential narrowing of research agendas, wherein work that aligns with deployable, monetizable outcomes is privileged over research that interrogates ethical trade-offs, social impacts, or long-term risks.

Moreover, the underreporting of funding sources further obscures the extent of corporate influence in shaping the knowledge landscape. This lack of transparency hinders the ability of external observers to critically assess how financial incentives may affect methodological choices, problem framing, or value prioritization in published research.

C. Implications for Research Diversity

The concentration of authorship and funding within a small set of elite and corporate institutions has significant implications for epistemic diversity and agenda-setting in machine learning. When research is disproportionately driven by entities with aligned economic or geopolitical interests, the scope of inquiry may contract, prioritizing performance improvements and technical optimization over critical reflection and societal responsiveness.

To foster a more inclusive and balanced research environment, increased attention must be given to diversifying funding pathways, promoting equitable authorship opportunities, and establishing norms around full disclosure of institutional and financial affiliations. Such measures are essential for ensuring that the field evolves not only in terms of technical sophistication, but also in its ethical maturity and global inclusivity.

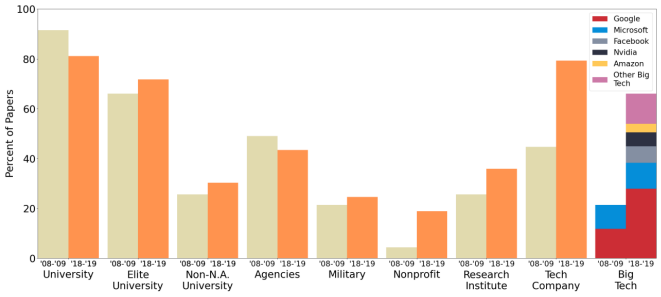

Fig. 3: Corporate and institutional ties of high-impact papers over time.

VI. Discussion

The findings of this study offer a substantive challenge to the prevailing narrative that machine learning research is inherently objective, apolitical, or universally beneficial. Instead, our analysis reveals a pronounced value orientation within highly cited ML literature—one that privileges technical metrics, institutional prestige, and deployability, often at the expense of ethical reflection, inclusivity, and social accountability.

One of the most striking patterns observed is the near-total omission of ethical considerations, even in papers addressing high-stakes domains such as healthcare, finance, or surveillance. Core ethical principles—fairness, transparency, accountability, autonomy, and human dignity—are rarely acknowledged, let alone meaningfully integrated into the research frameworks. This omission is not merely a matter of oversight but reflects an implicit prioritization of values like performance, novelty, and scalability, which are more easily

measured and often more directly aligned with corporate or academic incentives.

While performance metrics such as accuracy, F1 score, or inference time offer convenient benchmarks for progress, their dominance in justifying research contributions may obscure the contextual complexity of real-world deployment. A model achieving state-of-the-art results on benchmark datasets may still reproduce harmful biases, fail under domain shifts, or exacerbate existing inequalities in practical applications. Yet few papers engaged with these risks, and even fewer offered mechanisms for identifying or mitigating them. This signals a systemic undervaluation of harm-centered perspectives in the current ML publication ecosystem.

Furthermore, the growing centrality of elite institutions—both academic and corporate—in shaping the research discourse raises concerns about epistemic homogenization and agenda capture. As our institutional analysis indicates, the most visible and impactful research increasingly originates from a concentrated set of organizations with aligned strategic interests. While such institutions undoubtedly contribute valuable resources and infrastructure, their dominance risks marginalizing alternative research paradigms—particularly those emerging from underrepresented regions, community-based organizations, or disciplines outside the core ML community.

This concentration of authorship and funding may also shape the field's incentive structures. Research that advances the goals of commercial scalability, market integration, or technological innovation is often rewarded, while critical, interpretive, or justice-oriented work may struggle for recognition and support. As a result, the boundaries of "valuable" research are defined not merely by scientific rigor but by institutional alignment with dominant economic and political interests.

The absence of diverse voices—including those most affected by the deployment of ML systems—further compounds the problem. Without intentional inclusion of marginalized perspectives, the field risks reifying inequities under the guise of technical progress. Democratizing ML research will therefore require more than open-source code or broader dataset access; it will necessitate structural changes in publication practices, funding distribution, and community norms that currently privilege a narrow set of actors and values.

Ultimately, these findings underscore the need for a more reflective and inclusive machine learning research culture—one that critically interrogates its assumptions, explicitly acknowledges its limitations, and centers societal well-being as a primary criterion for progress. Future research must work toward rebalancing technical innovation with ethical deliberation, ensuring that the benefits of ML are both equitable and just.

VII. Conclusion

This study highlights the inherently normative character of contemporary machine learning research, directly challenging the widespread assumption of its neutrality or objectivity. Through a systematic content analysis of 100 highly cited papers published between 2008 and 2019 in premier ML venues, we uncovered a consistent and reinforcing pattern: values such as performance optimization, computational efficiency, and generalizability are overwhelmingly prioritized, while ethical, societal, and user-centered considerations are notably marginalized.

These dominant value preferences are not incidental. They are embedded within broader institutional structures—particularly the growing entanglement of elite academic institutions and corporate technology firms—which significantly influence what research is conducted, how it is evaluated, and whose interests it ultimately serves. The result is a research culture where social accountability, transparency, fairness, and harm reduction are treated as peripheral, rather than foundational, to technical progress.

Equally concerning is the near-total absence of rigorous engagement with the potential risks and unintended consequences of ML systems, even in domains where such harms are well-documented. This lack of critical discourse not only weakens the field's ability to safeguard against misuse but also erodes public trust in machine learning technologies.

To address these shortcomings, a redefinition of success within the ML research community is urgently needed. Progress must no longer be equated solely with surpassing benchmark datasets or publishing in prestigious venues. Instead, evaluation criteria should be broadened to include the social utility, contextual relevance, and ethical integrity of proposed methods. Incentive structures—such as peer review norms, funding priorities, and institutional recognition—must evolve accordingly.

Furthermore, cultivating an inclusive and pluralistic research environment will require intentional efforts to elevate voices and perspectives that have historically been excluded from the ML discourse. This includes researchers from underrepresented regions, disciplines focused on critical theory or social justice, and communities directly impacted by algorithmic systems.

In conclusion, steering machine learning toward a more socially responsible trajectory is not merely a matter of technical refinement. It is a political, institutional, and cultural undertaking—one that demands collective commitment to equity, accountability, and shared human flourishing. Only through such a deliberate shift can ML fulfill its potential as a force for inclusive and ethical innovation.

References

[1] S. H. Al Harbi, L. N. Tidjon, and F. Khomh, "Responsible Design Patterns for Machine Learning Pipelines," *arXiv preprint*, May 2023.

[2] T. LaCroix and S. J. D. Prince, "Deep Learning and Ethics," in *Understanding Deep Learning*, arXiv, May 2023.

[3] T. K. Gilbert, M. W. Brozek, and A. Brozek, "Beyond Bias and Compliance: Towards Individual Agency and Plurality of Ethics in AI," *arXiv preprint*, Feb. 2023.

[4] W. Ma and V. Valton, "Toward an Ethics of AI Belief," *arXiv preprint*, Apr. 2023.

[5] M. _*, "Mapping the Ethics of Generative AI: A Comprehensive Scoping Review," *Minds and Machines*, 2024.

[6] J. Juli, “Ethical Considerations in Artificial Intelligence and Machine Learning,” *EasyChair Preprint 12536*, Mar. 2024.
[7] P. Thakur *et al*., “Machine Learning and Data Ethics: a Design of Integrated Framework Towards Intelligent Decision Making,” *EasyChair Preprint 14102*, Jul. 2024.
[8] R. K. Paul and B. Sarkar, “Generative AI and Ethical Considerations for Trustworthy AI Implementation,” *IJAIML*, vol. 2, no. 1, 2023.
[9] S. N. Halder and S. Sarkar, “Ethical application of artificial intelligence and machine learning in research and education,” in *Academic Integrity and Innovation*, Prova Prakashani, 2024.
[10] N. Hussain and A. Lee, “Ethical Considerations in Artificial Intelligence and Machine Learning,” *J. Nonlinear Analysis and Optimization*, vol. 14, no. 1, 2023.
[11] Z. Hamid, S. Ajmal, and I. Torshin, “Ethical Considerations in AI and Machine Learning,” ResearchGate, Nov. 2023.
[12] S. T. Boppiniti, “Data Ethics in AI: Addressing Challenges in Machine Learning and Data Governance for Responsible Data Science,” *Int. Scientific J. for Res.*, vol.5, no.5, 2023.
[13] B. Memarian and T. Doleck, “Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review,” *Computers and Education: Artificial Intelligence*, 2023.
[14] S. Ruggieri *et al*., “Can We Trust Fair-AI?,” in *AAAI Conf. on Artificial Intelligence*, Jun. 2023.
[15] A. Castelnovo *et al*., “Fair Enough? A map of the current limitations of the requirements to have ‘fair’ algorithms,” Commun. ACM, 2023.
[16] M. R. Islam, “Generative AI, Cybersecurity, and Ethics,” John Wiley Sons, 2024.
[17] A. Gazis *et al*., “Organising AI for safety: Identifying structural vulnerabilities to guide the design of AI-enhanced socio-technical systems,” *Saf. Sci.*, Apr. 2025.
[18] “Transparency and the Black Box Problem: Why We Do Not Trust AI,” *Philosophy Technology*, Dec. 2025.
[19] B. Lund, “Standards, frameworks, and legislation for artificial intelligence (AI) transparency,” *AI and Ethics*, Jan. 2025.
[20] Y. Zhang, C. Dong, W. Guo, J. Dai, and Z. Zhao, “Systems theoretic accident model and process (STAMP): A literature review,” *Saf. Sci.*, 2022–2023.