How is the Socio-Demographic Background of Researchers in AI & ML Related to the Values reflected in their Research?

PAULA NAUTA, Graz University of Technology, Austria and Complexity Science Hub, Austria
FARIBA KARIMI, Graz University of Technology, Austria and Complexity Science Hub, Austria
ANA MARÍA JARAMILLO, Graz University of Technology, Austria and Complexity Science Hub, Austria

In this work we investigate the socio-demographic factors influencing the production of influential Artificial Intelligence (AI) and Machine Learning (ML) research. This work builds upon prior work, which identified a predominance of power-centralizing values and an underrepresentation of user rights and ethical principles in AI & ML publications [3]. Here we use the same dataset (*seed publications*) consisting of the most cited publications presented at top-tier conferences NeurIPS and ICML in four selected years: 2008, 2009, 2018, and 2019 and analyze whether the socio-demographic composition of authors influences the prevalence of these values.

An enriched dataset with all publications in the same conferences and years is constructed from open-access research platforms such as Semantic Scholar and Open Alex. Publications are identified as closely related to one of two groups derived from initial annotations in the Seed publications: (i) Moral Group and (ii) Non-Moral Group. This is achieved by computing Jaccard Similarity based on reference overlap between publication pairs in the enriched dataset and constructing a similarity-based network, followed by backbone (Figure 1**A**) and ego network extraction (Figure 1**B**). Diversity scores for research collaborations are calculated enabling a statistical analysis with the two groups of publications. Human validation assesses whether publications identified as similar to seed publications, also contain the same moral classification. Results reveal that although the developed method successfully constructs a similarity-based measure, it does not reliably infer shared moral values. Publications closely tied to a publication categorized as *moral* do not necessarily share the same values, despite having a high overlap based on their references. Additional results show that the diversity characteristics of research collaborations in both groups do not have a statistically significant relationship with the moral classification of publications. While some diversity is present, the results reveal a concentration of researchers from a few nationalities in the Global North (Figure 1**C**), an increasing share of corporate affiliations over time (Figure 1**D**), and a persistent dominance of elite institutions (Figure 1**E**), pointing to structural imbalances in the network.

Keywords: Diversity, Ethics in AI, Moral Values, Science of Science, Artificial Intelligence & Machine Learning Research, Network Science, Computational Social Science

Reference Format:
Paula Nauta, Fariba Karimi, and Ana María Jaramillo. 2025. How is the Socio-Demographic Background of Researchers in AI & ML Related to the Values reflected in their Research?. In *Proceedings of Fourth European Workshop on Algorithmic Fairness (EWAF'25).* Proceedings of Machine Learning Research, 6 pages.

Authors' Contact Information: Paula Nauta, nauta@tugraz.at, Graz University of Technology, Graz, Austria and Complexity Science Hub, Vienna, Austria; Fariba Karimi, karimi@tugraz.at, Graz University of Technology, Graz, Austria and Complexity Science Hub, Vienna, Austria; Ana María Jaramillo, jaramillo@csh.ac.at, Graz University of Technology, Graz, Austria and Complexity Science Hub, Vienna, Austria.

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

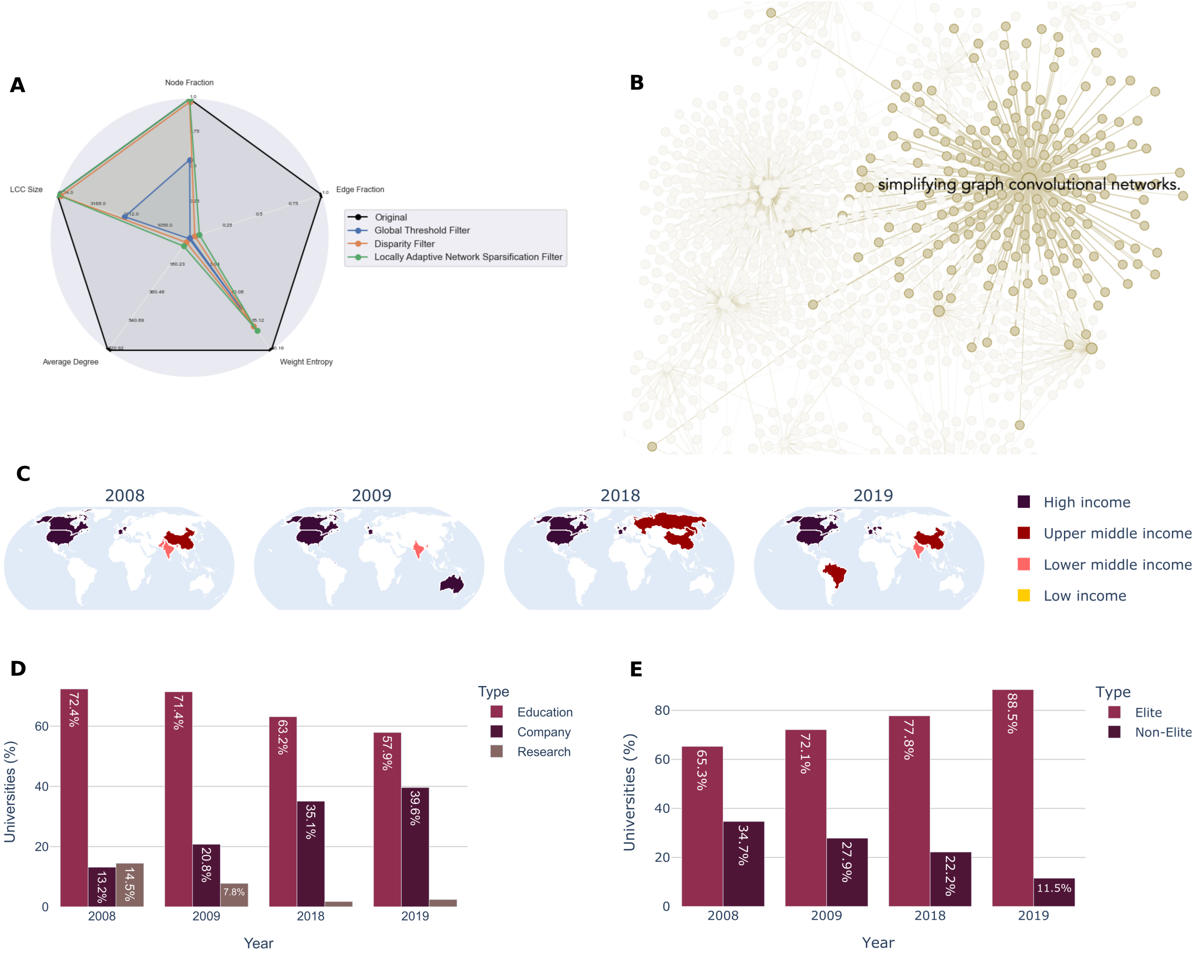

Fig. 1. **A** Radar Plot shows Backbone Retention techniques and their influence on diverse network metrics. **B** Extraction of a Moral Publication's Ego-Network. **C** Countries of Affiliations by Income Group. **D** Types of Institutions. **E** Disaggregation of the "Education" category from (D), comparing Elite vs. Non-Elite universities.

1 Introduction

With the rise of Large Language Models (LLMs) like OpenAI's ChatGPT in November 2022, AI and ML technologies have become central to discussions about their increasing role in various sectors of society [4]. While these advancements have undeniably transformed many aspects of life, their benefits are not universally distributed [10]. Over time, the disruptive potential and risks of these technologies have become increasingly evident as exemplified by [25]. The potential for disruption lies in the ability of ML algorithms to rapidly influence and distort societal structures, reshape decision-making processes, and automate tasks that were previously carried out by humans [22]. Marketed as objective and efficient, they often mask biases, and embed inequities into optimized processes [21]. For example, search engines have been shown to amplify harmful stereotypes, reflecting deep-seated societal biases that influence public and political discourse [13, 20]. In response, governments worldwide are establishing frameworks to regulate the development and use of AI & ML systems [7]. With this growing acknowledgment, it becomes paramount to examine the drivers and identify who is conducting the research behind these advancements.

Previous research qualitatively analyzed and identified the values present in highly influential AI & ML publications, using a ten-year time span to capture shifts in the value discourse [3]. These values refer to desirable attributes, both technical, such as performance and efficiency, and societal considerations, including fairness and ethics. In [3], the authors' findings revealed the predominance of power-centralizing values, while societal needs and ethical considerations were neglected and significantly underrepresented.

Scholars highlight how technical fields remain predominantly shaped by a narrow and non-diverse demographic, resulting in what is described as a "privilege hazard" [8]. This concept emphasizes how individuals in positions of privilege often fail to recognize systemic inequities and design systems that reflect only their own lived experiences. The researchers suggest that this demographic homogeneity limits the scope of diverse values in the scientific and technological priorities that influence society. Decisions about what is measured, how it is measured, and which problems are deemed important to solve are deeply influenced by the positionalities [1] of the individuals involved (ibid.). For example, historical exclusion of women and other marginalized groups from data collection and analysis has resulted in technological designs that fail to address women-specific needs, with significant consequences [23].

Yet previous literature reveals a positive impact of gender, ethnic, and institutional diversity on research outcomes, including innovation, interdisciplinary advancements, and, particularly for papers affiliated with top-tier institutions, higher citation rates [1, 2, 11, 15, 16, 18, 19, 27]. These patterns reflect a broader view of science as socially and institutionally embedded, where power dynamics and dominant interests shape what is considered valuable knowledge [5]. Building on this, we examine whether the values present in researchers' publications within AI & ML are related to their socio-demographic factors, with implications for the production and direction of scientific knowledge.

2 Methods

The analysis builds on a set of value-annotated *Seed publications* [3], which serve as ground truth. To distinguish between value orientations, these publications are first classified into the *Moral* group, if they include values tied to ethics or user rights, and into the *Non-Moral* group, if they reflect only technical values such as performance or efficiency. As publications reflecting ethical and user-focused principles are underrepresented in this seed set, an enlarged dataset is constructed containing all publications from the same conferences and years to enable meaningful statistical analysis.

The methodology used to examine the relationship between the socio-demographic composition of research teams and the values reflected in their publications consists of four main components: **(i)** Jaccard similarity is computed for all publication pairs in the population based on their reference overlap. The resulting similarity scores are used to construct a weighted co-citation network, where nodes represent publications and edge weights

[1] Positionalities refer to the social locations and identities of individuals, such as gender, race, ethnicity, class, sexuality, age, and ability, and how these intersect with power dynamics in shaping perspectives, experiences, and contributions within research or decision-making contexts [14].

reflect their reference similarity. **(ii)** A backbone extraction (Figure 1**A**) is applied to the resulting network to retain only statistically meaningful edges and reduce noise. This step ensures that high-similarity connections, especially those involving the seed publications, are preserved. **(iii)** Ego networks are extracted (Figure 1**B**) for each seed publication within both the *Moral* and *Non-Moral* groups. These ego networks represent the closest publications in terms of shared references. To validate whether these neighboring publications reflect the same value orientation as the seed, a research team with a quantitative background assessed a sample of extracted publications for the presence or absence of moral values. **(iv)** Finally, socio-demographic diversity is computed for each publication in the ego networks, considering the gender, nationality, and institutional affiliation of authors. We apply Stirling's diversity framework [24], which incorporates variety, balance, and disparity. In the case of affiliation diversity, disparity is operationalized using the GDP of each author's country of affiliation. These diversity scores are used in a Spearman correlation, logistic regression, and interaction effects analysis to test for associations between diversity and the publication's value orientation (*Moral* vs. *Non-Moral*).

3 Results & Discussion

The statistical analysis reveals similar distributional patterns between *Moral* and *Non-Moral* Ego Network publications. This is further supported by statistical tests, which show no significant relationship between any of the diversity metrics (gender, nationality, affiliation) and the ego network publications of either group. The initial hypothesis that collaborations with greater diversity emphasize more moral values contrasts with the overall homogeneity observed in the network, particularly in terms of gender and affiliations. The results suggest that diversity may not be a key indicator of moral values in the context of this research population. Instead, the increasing influence of corporate dominance (Figure 1**D**) seems to support the creation of standardizations and norms based on performance-driven science, which might facilitate the understanding across diverse national, corporate, and institutional cultures [6, 17]. While these frameworks support collaboration and operational efficiency, they may inadvertently suppress the emergence of diverse perspectives.

Literature showed that affiliation with corporate entities significantly boosts citation counts [1], which in turn might lead researchers to align and adapt with institutional values rather than pursue new ideas that are less accepted [12, 26]. Although this community exhibits some diversity, the findings align with the phenomenon of 'brain drain', which describes how global science and research talent increasingly concentrate in wealthier, resource-rich institutions and countries (Figure 1**C**), while the nationalities of the researchers are from lower-income countries [9]. Moreover, the concentration of influence among elite universities (Figure 1**E**), reflects broader societal patterns, where resources and opportunities disproportionately accumulate within a few dominant actors, suggesting that research priorities are shaped mainly by their agendas. Regardless of their background, researchers consequently align with these paradigms to succeed and gain visibility. While there may be some degree of diversity within the field, the prevailing system continues to limit and constrain the development of alternative perspectives, thereby maintaining the status quo.

4 Limitations & Future Work

The methodology employed to identify publications reflecting particular value orientations requires further refinement. A central limitation, highlighted during the human evaluation phase, concerns the difficulty of reliably identifying values in scientific publications. The likelihood of a value being explicitly mentioned is shaped by its contextual salience and degree of institutionalization within the field. For instance, widely recognized values such as *fairness* or *privacy* may be explicitly stated, whereas less prominent values such as *beneficence* or *non-maleficence* are often implicit or require substantial domain knowledge to be identified. To this end, we propose the development of methods that leverage large language models (LLMs) to allow large-scale identification of both explicit and implicit value expressions in scientific texts.

Additionally, future research could investigate whether certain researchers or institutions exhibit consistent commitments to particular values, and whether these patterns persist across time or emerge intermittently. A longitudinal and actor-focused analysis may offer further insights into the social organization of value emphasis in AI and ML research.

Finally, there is a need to expand how affiliation diversity is operationalized. While GDP-based measures capture macroeconomic disparities, they do not account for epistemic or disciplinary convergence. Institutions from economically distinct countries may still exhibit substantial topical homogeneity if embedded within similar scientific paradigms. Future work should incorporate indicators that reflect disciplinary orientation and institutional type (e.g., academic, corporate, non-profit, governmental) to assess the diversity of research contexts more precisely.

Acknowledgments

We thank Abeba Birhane and her collaborators for their foundational work in AI ethics and sociotechnical critique, which significantly inspired the direction and framing of this work. We are also grateful to the anonymous reviewers for their constructive feedback. Finally, we thank the EWAF conference organizers for the opportunity to present and publish this work.

References

[1] Jonathan Adams. 2013. The fourth age of research. *Nature* 497, 7451 (May 2013), 557–560. https://doi.org/10.1038/497557a

[2] Bedoor K. AlShebli, Talal Rahwan, and Wei Lee Woon. 2018. The preeminence of ethnic diversity in scientific collaboration. *Nature Communications* 9, 1 (Dec. 2018), 5163. https://doi.org/10.1038/s41467-018-07634-8

[3] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The Values Encoded in Machine Learning Research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 173–184. https://doi.org/10.1145/3531146.3533083

[4] Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. 2023. Science in the age of large language models. *Nature Reviews Physics* 5, 5 (2023), 277–280. https://doi.org/10.1038/s42254-023-00581-4

[5] Pierre Bourdieu. 1975. The specificity of the scientific field and the social conditions of the progress of reason. *Social Science Information* 14, 6 (Dec. 1975), 19–47. https://doi.org/10.1177/053901847501400602

[6] Geoffrey Bowker and Susan Leigh Star. 1999. Sorting things out. *Classification and its consequences* 4 (1999).

[7] Angela Daly, Thilo Hagendorff, Hui Li, Monique Mann, Vidushi Marda, Ben Wagner, Wei Wang, and Saskia Witteborn. 2019. Artificial Intelligence, Governance and Ethics: Global Perspectives. *SSRN Electronic Journal* (2019). https://doi.org/10.2139/ssrn.3414805

[8] Catherine D'ignazio and Lauren F Klein. 2020. *Data feminism*. MIT press.

[9] Frédéric Docquier and Hillel Rapoport. 2012. Globalization, brain drain, and development. *Journal of economic literature* 50, 3 (2012), 681–730.

[10] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

[11] Bas Hofstra, Vivek V Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A McFarland. 2020. The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences* 117, 17 (2020), 9284–9291.

[12] Bas Hofstra, Daniel A. McFarland, Sanne Smith, and David Jurgens. 2022. Diversifying the Professoriate. *Socius* 8 (Jan. 2022), 23780231221085118. https://doi.org/10.1177/23780231221085118

[13] Masabah Bint E. Islam, Muhammad Haseeb, Hina Batool, Nasir Ahtasham, and Zia Muhammad. 2024. AI Threats to Politics, Elections, and Democracy: A Blockchain-Based Deepfake Authenticity Verification Framework. *Blockchains* 2, 4 (Nov. 2024), 458–481. https://doi.org/10.3390/blockchains2040020

[14] Kelly F Jackson, Sara Goodkind, Mery Diaz, Sharvari Karandikar, Ramona Beltrán, Mimi E Kim, Jennifer R Zelnick, Margaret F Gibson, Sarah Mountz, Gina E Miranda Samuels, et al. 2024. Positionality in critical feminist scholarship: situating social locations and power within knowledge production. , 5–11 pages.

[15] Benjamin F. Jones, Stefan Wuchty, and Brian Uzzi. 2008. Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science. *Science* 322, 5905 (Nov. 2008), 1259–1262. https://doi.org/10.1126/science.1158357 Publisher: American Association for the Advancement of Science.

[16] Aparna Joshi. 2014. By whom and when is women's expertise recognized? The interactive effects of gender and education in science and engineering teams. *Administrative Science Quarterly* 59, 2 (2014), 202–239.

[17] Michèle Lamont and Virág Molnár. 2002. The study of boundaries in the social sciences. *Annual review of sociology* 28, 1 (2002), 167–195.

[18] Mathias Wullum Nielsen, Sharla Alegria, Love Börjeson, Henry Etzkowitz, Holly J Falk-Krzesinski, Aparna Joshi, Erin Leahey, Laurel Smith-Doerr, Anita Williams Woolley, and Londa Schiebinger. 2017. Gender diversity leads to better science. *Proceedings of the National Academy of Sciences* 114, 8 (2017), 1740–1742.

[19] Mathias Wullum Nielsen, Jens Peter Andersen, Londa Schiebinger, and Jesper W Schneider. 2017. One and a half million medical papers reveal a link between author gender and attention to gender and sex analysis. *Nature human behaviour* 1, 11 (2017), 791–796.

[20] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press. https://doi.org/10.18574/nyu/9781479833641.001.0001

[21] Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, USA.

[22] Matteo Pasquinelli. 2023. *The Eye of the Master: A Social History of Artificial Intelligence*. Verso Books.

[23] Caroline Criado Perez. 2019. *Invisible Women: the Sunday Times number one bestseller exposing the gender bias women face every day*. Random House.

[24] Andy Stirling. 2007. A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface* 4, 15 (Aug. 2007), 707–719. https://doi.org/10.1098/rsif.2007.0213

[25] Ibo van de Poel, Lily Eva Frank, Julia Hermann, Jeroen Hopster, Dominic Lenzi, Sven Nyholm, Behnam Taebi, and Elena Ziliotti. 2023. *Ethics of Socially Disruptive Technologies: An Introduction*. Open Book Publishers. https://doi.org/10.11647/obp.0366

[26] Caroline S. Wagner, J. David Roessner, Kamau Bobb, Julie Thompson Klein, Kevin W. Boyack, Joann Keyton, Ismael Rafols, and Katy Börner. 2011. Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics* 5, 1 (Jan. 2011), 14–26. https://doi.org/10.1016/j.joi.2010.06.004

[27] Yang Yang, Tanya Y. Tian, Teresa K. Woodruff, Benjamin F. Jones, and Brian Uzzi. 2022. Gender-diverse teams produce more novel and higher-impact scientific ideas. *Proceedings of the National Academy of Sciences* 119, 36 (Sept. 2022), e2200841119. https://doi.org/10.1073/pnas.2200841119