What Impact Does Big Tech Funding Have on AI Research? A Scholarly Document Analysis

Max Martin Gnewuch
maxmartin.gnewuch@stud.uni-goettingen.de
Georg-August-University of Göttingen
Göttingen, Lower Saxony, Germany

ABSTRACT

For more than four decades, artificial intelligence (AI) research has thrived at the crossroads of academia and industry. However, the balance of influence is increasingly shifting toward industry, which dominates key components of modern AI research: computational resources, large datasets, and highly skilled researchers. This dominance is reshaping research outputs, with industry becoming more influential in academic publications. Yet, little empirical work examines the influence of industry on AI research (past or present). In this thesis, I quantify industry's presence in top-tier AI conferences and its citational influence on research trajectories. I analyzed ~57.3K AI papers, ~3.3M citations from AI papers to other papers, and ~6M citations from other papers to AI papers. I find a striking paradox: although the share of industry-funded research in top AI conferences has declined by 37 % between 2020 and 2023, its citational influence continues to grow. More than half (54 %) of industry-funded papers published between 2018 and 2023 have a high citational impact, compared to just 3 % of non-industry-funded and 2 % of non-funded papers. In addition, I find that industry-funded papers are insular - citing increasingly more industry-funded papers while contributing fewer papers that bridge diverse funding sources. Furthermore, I show that the temporal citation diversity of industry-funded papers have markedly declined, with both median age and age diversity at all-time lows. My findings raise questions about the scientific community's engagement with cross-disciplinary and temporally diverse literature, particularly in the context of industry-funded research. All data and code used in this thesis are publicly available.[1]

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Social and professional topics** → *Surveillance*; Governmental regulations; **Computing education**.

KEYWORDS

Computer Science, Scientometrics, Research Trends, Artificial Intelligence, Research Funding, Big Tech, Monopolisation, Conflict of Interest, Echo Chamber, Bias, Ethical AI, Fairness, Diversity, Transparency

1 INTRODUCTION

Artificial Intelligence (AI) has evolved into a "general purpose technology" comparable to the steam engine, electrification, and the Internet, offering transformative opportunities across various industries [Ahmed and Wahed 2020; Cockburn et al. 2018; Verdegem 2024]. This evolution holds the potential to reshape the social landscape by affecting job development, influencing hiring decisions, and addressing global challenges like climate change [Dobbe and Whittaker 2019]. However, growing concerns about biases and fairness in AI-enabled technologies persist [Ahmed et al. 2023; Bolukbasi et al. 2016; Nadeem et al. 2020; Righetti et al. 2019; Verdegem 2024]. Engaging a diverse group of researchers in AI development is crucial for mitigating biases and promoting fairness in AI practices [Kuhlman et al. 2020; West et al. 2019]. Thus, there is a broad consensus that the development of AI systems should be inclusive, ensuring that their benefits are widely accessible rather than limited to a privileged few [Ahmed and Wahed 2020].

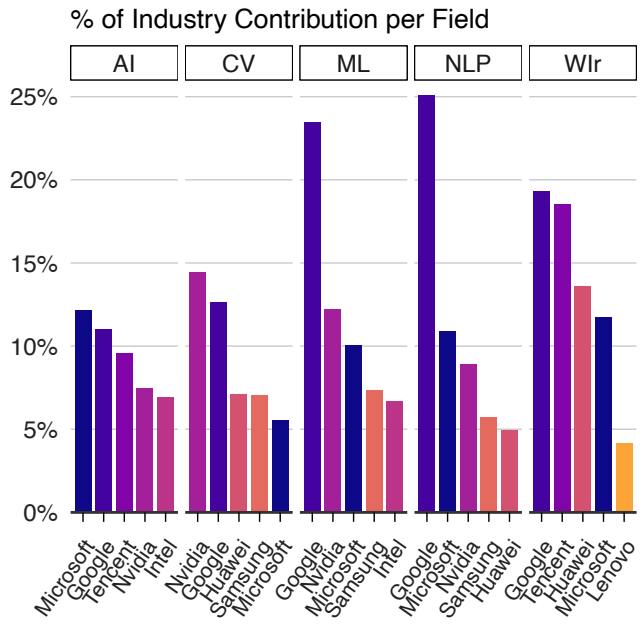


Figure 1: The contribution share of the top five industry funders in AI research across key domains from 2018 to 2023: Artificial Intelligence (AI), Computer Vision (CV), Machine Learning (ML), Natural Language Processing (NLP), and Web & Information Retrieval (WIr).

The dominance of Big Tech. However, the industrial landscape of AI is dominated by a small number of extremely powerful companies (see Figure 1), often referred to as *Big Tech*[2] (e.g., Meta, Google, Amazon, Nvidia, Microsoft, Baidu, Tencent, etc.) [Abdalla et al. 2023b; Ahmed and Wahed 2020; Montes and Goertzel 2019;

[1] https://github.com/Peerzival/impact-big-tech-funding

[2] The terms *Big Tech* and *industry* are used interchangeably.

Riedl 2020; Verdegem 2024]. These companies have access to three key resources essential for modern AI research: large datasets, the computational power necessary to run and train advanced machine/deep learning models, and access to a highly skilled AI workforce [Ahmed et al. 2023; Montes and Goertzel 2019; Riedl 2020; Verdegem 2024]. This oligopolistic/monopolistic control over these critical resources grants a disproportionate amount of power to a small number of corporations [Ahmed et al. 2023; Montes and Goertzel 2019], who ultimately shape what we examine (and do not examine) about AI and the business surrounding it [Whittaker 2021]. As a result, there is growing concern about scientific independence and the concentration of power in AI research [Abdalla and Abdalla 2021; Ahmed et al. 2023; Whittaker 2021].

Tensions between commercial and public interests. Whittaker [2021] argued that the close ties between the AI research community and industry do not necessarily compromise researchers in the domain of AI. However, these connections do mean that the questions and incentives shaping the field are not always within the control of individual researchers. The trajectory of the field — including which questions are deemed worth pursuing and which answers lead to grants, awards, and tenure — is disproportionately influenced by the corporate emphasis on resource-intensive AI and the tech industry's incentives.

The increasing investment from industry in AI research does not diminish the potential for significant societal benefits. Yet, commercial motives often drive companies to prioritize profit-oriented topics. While such incentives can sometimes align with the public interest, resulting in beneficial tools, hardware, or software, this is not always the case [Ahmed et al. 2023]. The lack of public-minded alternatives may lead to a scenario similar to that of the pharmaceutical industry, where investments tend to overlook the needs of lower-income groups [Ahmed et al. 2023; Trouiller et al. 2002].

This thesis. Does not aim to debate the advantages or drawbacks of the industry's growing presence in AI. Instead, it offers a systematic, quantitative evaluation of industry's influence within the AI research ecosystem. Using the Scopus database[3], I compiled a new dataset of metadata associated with ~57.3K AI papers published at ten top-tier conferences from 2018 to 2023, along with ~3.3M citations from AI papers and ~6M citations to AI papers. Each citation is detailed by year of publication, funding agency (if any), and field of study for both the papers *cited* by AI research and those *citing* AI research. Analyzing these connections allows for a in-depth exploration of five critical areas:

(1) *Evolution of industry presence in AI*: Analysis of trends in industry-funded research over time.
(2) *Engagement of the AI community*: Examination of the AI community's interaction with industry-funded research and how this engagement has evolved.
(3) *Insularity of industry-funded research by funding type*: Investigation of whether industry-funded research predominantly builds on its own findings or follows the principles of responsible research, aligning with responsible research principles.
(4) *Insularity of industry-funded research by field*: Given the broad social impact of AI, particularly through subfields such as NLP and ML, it is important to assess whether Big Tech firms, such as Google, Microsoft, and Tencent, follow their societal responsibilities. Specifically, this analysis examines whether these companies research engages with a wide swathe of fields, particularly in areas relevant to the societal impact of AI.
(5) *Insularity of industry-funded research by citation age*: Sustainable and responsible research builds upon broad set of literature, spanning from various fields and periods. However, recent studies have shown a trend towards research insularity, characterized by a concerning reliance on recent publications at the expense of foundational work [Bollmann and Elliott 2020; Nguyen and Eger 2024; Singh et al. 2023; Verstak et al. 2014; Wahle et al. 2024]. In the context of AI's societal impact, such a shift in declining engagement with older research could undermine the long-term sustainability of the field, research in general, and, ultimately, society. Therefore, this thesis examines trends in citation age across funding types over time, and explores the extent to which industry-funded research adheres to principles of responsible and inclusive research.

In summary, my contributions are three-fold: I (1) introduce a novel methodology for identifying corporate involvement in research, (2) compiled a new open dataset of ~57.3K AI papers, and (3) provide a detailed analysis of how deeply the AI community interacts with industry-funded research. It further examines whether industry-funded work engages with past literature and fosters interdisciplinary collaboration — both critical for responsible AI research.

By drawing upon the compiled dataset, I demonstrate that industry presence in the top AI conferences decreased by 37 % between 2020 and 2023. However, the AI community's engagement with industry-funded research has intensified. For example, in 2018, 9 %, or every ~12th citation from non-industry-funded or non-funded papers were directed toward industry-funded work. By 2023, this figure declined to every ~8th citation (12 %). This intensified engagement resulted in 54 % of industry-funded papers from 2018-2023 had high citational impact, compared to just 3 % of non-industry-funded and 2 % of non-funded papers.

These findings raises a cause for concern about the integrity and impartiality of current AI research. Some actions to address these challenges could include a stricter code of ethics and operation for AI-related research, enhanced regulation of Big Tech's role in research funding, and increased public resources, such as a public research cloud, public data sets, salaries, and research funding.

2 RELATED WORK

The field of scientometrics, particularly the analysis of citation patterns, has been a prominent area. Tracing back to the mid-20th century [de Solla Price 1962], citations and their networks have been studied from several perspectives, including: author location Rungta et al. [2022], affiliation [Abdalla et al. 2023b; Sin 2011], reputation [Collet et al. 2014], as well as demographic attributes such as gender, race and age [Abdalla et al. 2023a; Ayres and Vars

[3]The in-house Scopus database maintained by the German Competence Centre for Bibliometrics (Scopus-KB), 2024 version.

2000; Chatterjee and Werner 2021; Llorens et al. 2021; Mohammad 2020c]. Other perspectives include paper length [Falagas et al. 2013] and quality [Buela-Casal and Zych 2010], field of study [Costas et al. 2009], publication language [Lira et al. 2013] and venue [Wahle et al. 2022b], self-citation [Della Sala and Brooks 2008], plagiarism [Gipp and Meuschke 2011; Wahle et al. 2023a, 2022a] and institutional diversity [Abdalla et al. 2023b].

Despite AI's pivotal role in contemporary science, comprehensive scientometric studies of the entire field remain sparse. Most analysis have centered on Natural Language Processing (NLP), a prominent AI subfield. Many researchers have shared open-sourced datasets that can be used to study the growth and change in NLP [Mariani et al. 2019; Mohammad 2020a; Wahle et al. 2023b, 2022b].

Recent studies have begun to quantify the presence of industry in AI research. Ahmed et al. [2023] highlights that while AI originated within academia, industry now dominates its practical application, further development, and broad rollout. Similarly, Sevilla et al. [2022] analyzed trends in computational resources, revealing that industry's contribution to the largest AI models has gone from 11 % in 2010 to 96 % by 2021. Ahmed and Wahed [2020] created a dataset of 171K computer science conference papers to study industry participation rates over time. They found an upward trend in industry participation, largely limited to collaborations between major corporations and top-ranked research institutions (ranked 1-50 nn QS World University Rankings[4]). Moreover, Klinger et al. [2022] explored the thematic diversity of AI research, comparing academic and industrial contributions. They concluded that thematic diversity has stagnated, with industry-driven research being less diverse yet more influential than research in academia.

Building upon these existing studies, my thesis examines the (temporal) citational impact of industry-funded papers on AI research by analyzing a dataset covering papers that are published in the top-tier AI conferences (published between 2018 and 2023). Inspired by Abdalla et al. [2023b]; Singh et al. [2023]; Wahle et al. [2023b], this thesis delves into patterns of citation amnesia within AI, extending Singh et al. [2023]'s findings on citation amnesia in NLP. It addresses the trends of declining interdisciplinary engagement noted by Wahle et al. [2023b] and situates the observations of Abdalla et al. [2023b] on industry's influence within the broader AI context. My thesis takes a deep dive into eight novel research questions, notably around industry funding trends in AI (Q1), citation behaviors related to funding types (Q2, Q3, Q6), influence and citational impact of industry-funded research (Q4, Q5), and citation ages of industry-funded papers (Q7, Q8).

3 METHODOLOGY

To obtain robust data on industry presence in AI research, I employed the Scopus database for its extensive and inclusive content coverage [Pranckutė 2021]. Spanning publications from 1902 to 2024, Scopus includes ~69.5M papers and ~2.2B citations, providing detailed metadata on author affiliations and citation relationships. Additionally, Scopus includes funding information for ~21M papers. However, since this funding data relies on the information provided in the publication acknowledgments [Liu 2020], the identified industry funding affiliations reflect a lower bound. An overview of the key dataset statistics is provided in Table 1.

Table 1: Overall dataset statistics.
[†]Lower bound. [*]Sum of articles in Table 2.

Time range	1902–2024
#papers	69 491 766
#funded papers[†]	21 047 938
#citations	2 199 264 185
#papers AI[*]	114 090
#funded papers AI[†]	45 893
#out-citations from AI	3 308 618
#in-citations to AI	6 012 570

For the analysis of AI subfields, I selected five key domains: *Artificial Intelligence* (AI), *Computer Vision* (CV), *Machine Learning* (ML), *Natural Language Processing* (NLP), and *Web & Information Retrieval* (WIr), based on csranking.org[5]. From these domains, I included two top conferences per domain according to csranking.org and based on their h5-index rankings. Out of the ten conferences, eight were successfully matched, while the two unmatched conferences were replaced by those with the third-highest h5-index in their respective fields. The final list of top-tier AI conferences contains:

- Advancement of Artificial Intelligence (AAAI)
- International Joint Conference on Artificial Intelligence (IJCAI)
- Conference on Computer Vision and Pattern Recognition (CVPR)
- International Conference on Computer Vision (ICCV)
- International Conference on Machine Learning (ICML)
- International Conference on Learning Representations (ICLR)
- Association for Computational Linguistics (ACL)
- Empirical Methods in Natural Language Processing (EMNLP)
- International Conference on Web Search and Data Mining (WSDM)
- Conference on Research and Development in Information Retrieval (SIGIR)

Table 2 provides details of the selected top-tier AI conferences.

The introduction of the transformer network architecture in 2017 [Vaswani 2017], an algorithmic technique developed by Google for training neural networks, marked a significant shift in the development of language models. This advancement led to the emergence of Bidirectional Encoder Representations from Transformers (BERT) and similar models, accelerating the development of large-scale pretrained models [Devlin et al. 2019] and contributing to the monopolization of AI development in the digital economy [Luitse and Denkena 2021].

To capture the marked impact of transformer-based models on AI research, I focus on the period from 2018 to the present. This timeframe aligns with the introduction and widespread adoption of

[4]www.topuniversities.com/qs-world-university-rankings

[5]https://csrankings.org/#/index?ai&vision&mlmining&nlp&inforet&world

Table 2: The selected top-tier AI conferences are ordered by field and decreasing h5-index.
***Replacements with the third-highest h5-index.**

Field	Conference	Number of articles	h5-index (↓)
AI	AAAI	9696	212
	IJCAI	10 375	133
CV	CVPR	22 866	422
	ICCV*	11 691	291
ML	ICLR	4407	303
	ICML*	19 719	288
NLP	ACL	21 654	192
	EMNLP	6950	176
WIr	SIGIR	5145	90
	WSDM*	1587	77

these technologies. However, the Scopus data on the selected conferences only extends to 2023, so I focus on AI publications between 2018 and 2023. This range is both practical and comprehensive for analyzing funding trends driven by industry during the era of large-scale pretrained models. The final dataset contains 57 319 papers. Below, I outline my approach for collecting and processing this data to identify industry funding.

The source code to process my data and reproduce the experiments is available on GitHub (for research purposes only):

https://github.com/Peerzival/impact-big-tech-funding

3.1 Data Collection

To define the scope of Big Tech for the experiments, I constructed a citation graph that extends two levels deep from the conference papers listed in top-tier AI conferences. Figure 2 provides an example of this citation graph. The graph originates from papers presented at these top-tier conferences (root level) and expands through the outgoing citations of these papers (level 1). The final level (level 2) includes papers cited by those in the first level. From this graph, I identified corporate funders (CF) contributing to AI research.

This two-level expansion captures both direct and indirect contributors to the field, providing a holistic view of the industry influence on AI research. Many foundational contributions originate from companies that may not publish directly in these conferences, but are highly cited for their impactful work. For example, organizations such as OpenAI, Hugging Face, and Mistral play a pivotal role in advancing the field, even though they may not always present at these conferences.

This citation-based methodology addresses the limitations of previous approaches, such as that of Abdalla et al. [2023b], which emphasized market capitalization by focusing on publicly traded companies listed on the New York Stock Exchange. By including influential publicly and non-publicly traded companies, my approach ensures a more accurate representation of the entities shaping the AI research landscape.

3.2 Data Processing

The process to reproduce my list of CFs can be described in five steps:

(1) I extracted the names of funding agencies and the frequency of their occurrences from papers in the citation graph.
(2) For agencies with up to ten occurrences, I manually examined them to determine whether they were industry-related. This step is referred to as **manual analysis**. A funding agency was classified as CF if it was neither a public nor a non-profit organization.
(3) I standardized the names of the CFs to account for variations in company names and alternative descriptions of the same organization (e.g., Amazon, Amazon Web Services, Amazon Research).
(4) I used fuzzy matching (Appendix A.1) to search for CFs in the remaining set of agencies with fewer than ten occurrences. As a proxy for industry affiliation, I used the 216 standardized agencies from the third step. This step is referred to as **automatic analysis**.
(5) Finally, I merged the results from both the manual and automatic analyses into a comprehensive list of CFs. Table 3 provides an overview of key dataset statistics.

In total, I processed 78 333 funding agencies and identified 3136 as industry-related. Further details on the standardization process can be found in Appendix A.1.

Table 3: Industry-related funding agency dataset statistics.
***Sum of manual and automatic analysis.**

Attribute	Amount
Total funding agencies	78 333
Funding agencies up to ten occurrences	4206
Total industry affiliations*	3136
Industry affiliations (manual analysis)	382
Industry affiliations after standardization	216
Industry affiliations (automatic analysis)	2754

4 EXPERIMENTS

The integration of manually and automatically extracted CFs with the Scopus database allows an in-depth analysis of industry involvement in AI research across different subfields. In the following, I address eight pivotal questions regarding the role of industry in various areas of AI research.

***Q1.* How large is the industry funding in AI? How does this number vary by research field, such as AI, CV, ML, NLP, and WIr? Has this number stayed roughly the same or has it changed markedly over the years?**

Ans. To quantify industry funding, I cross-referenced funding agencies of AI papers with the CF dataset. The *percentage of industry-funded papers* (*PIFP*) in a given year y is calculated as:

$$PIFP(y) = \sum_{\forall f_i \in F} \frac{IF(f_i, y)}{P(f_i, y)} \cdot 100 \tag{1}$$

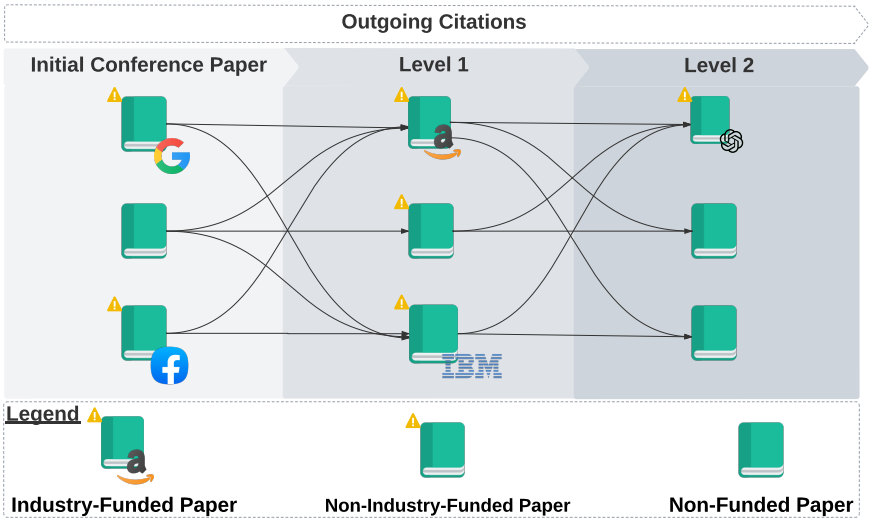


Figure 2: Example citation graph to identify funding agencies for the search scope.

here $IF(f_i, y)$ represents the number of industry-funded papers in field f_i in year y, and $P(f_i, y)$ is the total number of papers in that field. F is the set of all subfields in AI.

To gain finer granularity, I also compute the *field-specific industry funding percentage* (*FIFP*) for each year:

$$FIFP(x, y) = \frac{IF(x, y)}{\sum_{\forall f_i \in F} P(f_i, y)} \cdot 100 \tag{2}$$

This metric isolates industry funding trends within individual subfields.

For example, in 2020, if the total number of papers across all fields is 5000 and 300 papers in field x were industry-funded, the *FIFP* for x in 2020 would be $FIFP(x, 2020) = \frac{300}{5000} \cdot 100 = 6\%$.

Results. Figure 5 tracks the evolution of industry involvement in AI research, with increasing granularity from left to right. The long-term perspective on industry-funded AI research is shown in Figure 5(a), while Figure 5(b) narrows the focus to the selected time frame. Figure 12 in Appendix details industry funding distribution across the five AI subfields.

The proportion of industry-funded AI research has increased markedly, from 0.6 % in 1998 to 7% in 2023 (see Figure 5(a)). A sharp growth occurred between 2016 (4 %) and 2020 (11 %), reflecting a 180 % increase. However, Figure 5(b) shows a decline post-2020, with the share of industry-funded papers dropped from 11 % in 2020 to 7 % in 2023. The subfields most affected by this decline were AI and ML, with ML experiencing a dramatic 100 % reduction in industry involvement by 2023 (see Figure 12). In contrast, CV, NLP, and WIr have seen growth in industry presence. NLP remained the leading field in industry-funded research, while CV moved from fourth to second place by 2023. WIr, the only subfield showing consistent growth, surpassed ML in industry presence by 2023. Interestingly, ML and CV attract less industry funding than NLP, possibly due to the increasing demand for language-based technologies in virtual assistants, chatbots, and translation services.

Between 2018 and 2023, 9 % of all papers published at top-tier AI related conferences were funded by industry (see Figure 3). Figure 4 shows the percentage of industry-funded papers in an AI subfield out of all industry-funded papers. NLP (33 %) and CV (27 %) had the largest share of industry-funded publications, representing key fields of industry interest. Observe that despite its growth, WIr (3 %) lagged behind in attracting industry investment. An alternative perspective on the distribution of industry-funded publications across AI subfields is provided in Figure 13 within the Appendix.

Overall, 60 % of papers were funded, with industry providing less funding than other public sources (see Figure 3). Papers funded by non-profit or public organisations make up the largest share across all subfields, except in ML. 54 % of ML papers between 2018 and 2023 were not funded, while all other fields had a funding percentage above 50 %. The greatest concentration of industry-funded research occurred in NLP (11 %) and ML (10 %).

Discussion. The results align with Ahmed et al. [2023]'s observations of increasing industry presence in AI between 2016 and 2020, though my analysis reveals lower percentage values due to its focus on funding rather than institutional affiliations. Funding directly shapes research directions [Thelwall et al. 2023], serving as a more precise indicator of industry influence than affiliations alone.

The decline in industry funding post-2020 does not necessarily imply a reduced industry presence in AI research. Instead, preprint platforms such as arXiv show an exponential increase in AI-related papers [Krenn et al. 2023; Maslej et al. 2023]. This suggests that industry shifts away from traditional conference publications in favor of less costly and time-consuming alternatives like preprint servers (e.g., BERT [Devlin et al. 2019]) or corporate websites (e.g., ChatGPT[6]). The industry's ability to conduct exclusive research - owing to their access to critical resources - explains why these papers remain highly relevant, despite avoiding peer-reviewed venues.

[6]https://chat.openai.com

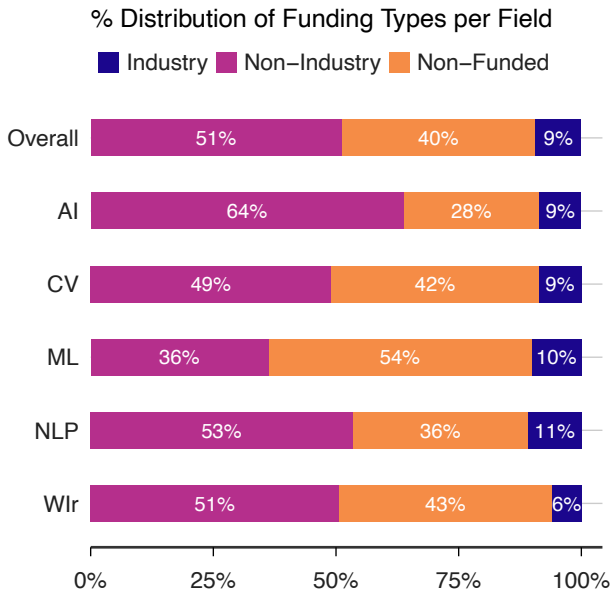


Figure 3: The percentage distribution of funding types (*industry*-funded paper, *non-industry*-funded paper, *non-funded* paper), overall and split by AI subfields.

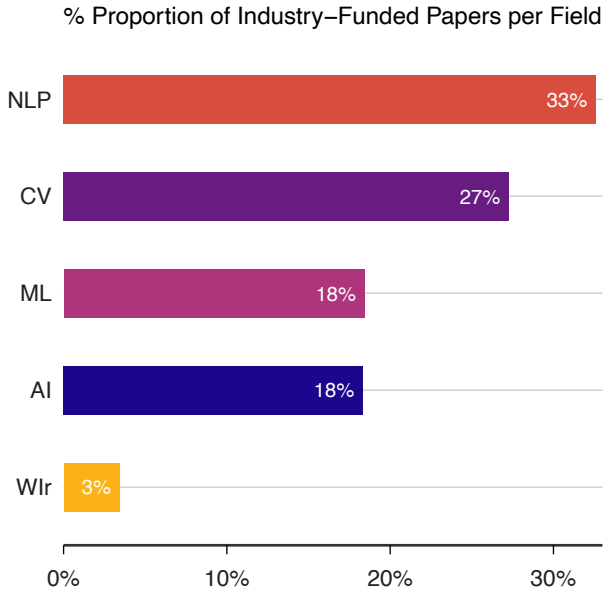


Figure 4: The percentage share of industry-funded papers in AI subfields (2018-2023), in descending order.

The COVID-19 pandemic further accelerated the use of preprint servers as a rapid publication medium [Smart 2022], while macroeconomic pressures such as supply chain disruptions, plummeting revenues, and rampant inflation have driven companies to cut costs [Morgan 2023], possibly influencing their reduced investment in academic conference publications. The potential trend of industry migration from publishing at top AI conferences to publishing on preprint servers and own websites is alarming, as the AI research community depends on industry funding and contributions. Such a shift may reinforce the perception that preprint publications alone constitutes good science, potentially devaluing peer review standards. This trend introduces several challenges, including limited time for researchers to filter quality work, promotion of bad science due to the lack of peer review standards, and decreased error correction via errata and retractions [Smart 2022].

Between 2018 and 2023, industry funding concentrated on NLP and CV, with limited investment in ML and AI (focus on symbolic techniques). This funding pattern reflects an industry preference for AI applications enabled by deep learning, alongside the development of infrastructure essential for scalable, safe deep learning research Klinger et al. [2022]. Non-deep learning AI methods and broader AI applications remain areas of limited industry interest. Conversely, WIr has shown steady funding growth, tied to its central role in developing search engines and recommendation systems, key AI application fields for technology companies.

The high proportion of funded papers - particularly in AI (73 %) and NLP (64 %) - highlights the financial demands of cutting-edge AI research, exceeding what unfunded individuals or groups can manage. Public and non-profit institutions are the primary supporters of AI research, although the rising share of industry funding in certain fields reflects growing corporate interest in the field's commercial applications. However, as Scopus bases funding information on paper acknowledgments, the true industry presence may be underreported, as disclosures are sometimes omitted [Wang and Shapira 2015]. Furthermore, funding in AI research extends beyond the flow of money from a funder to a recipient. The applied models created by industry are often those that push the boundaries of basic research [Ahmed et al. 2023]. Thus, funding includes access to models, datasets, computational power, and specialised expertise [Verdegem 2024].

ML remains an outlier, with most of its papers unfunded, likely a consequence of the field's theoretical and algorithmic focus, which requires fewer resources compared to experimental work.

***Q2.* Which funding types do AI papers cite more prominently? How has this citation behaviour changed over time?**

Ans. As we know from **Q1.**, 9 % of all papers published between 2018 and 2023 were funded by industry. The key question is whether industry-funded research attracts more citations than non-industry-funded and non-funded papers. Industry ownership of key resources [Ahmed and Wahed 2020; Ahmed et al. 2023; Verdegem 2024] may increase the visibility of these studies, even if their quantity is low. To answer this question, I take into account the different sizes of the funding types (industry, non-industry, and non-funded) and normalise the citation data with respect to the size of each funding type. I introduce a new metric, called the *Citation Preference Ratio* (*CPR*), which measures whether a paper with a certain funding type is cited more or less frequently than expected, based on its availability. In simple terms, a higher CPR indicates that a paper with a certain funding type is cited more often than its size in the literature would suggest, implying a positive citation preference for that type of funding. This metric helps to understand systemic

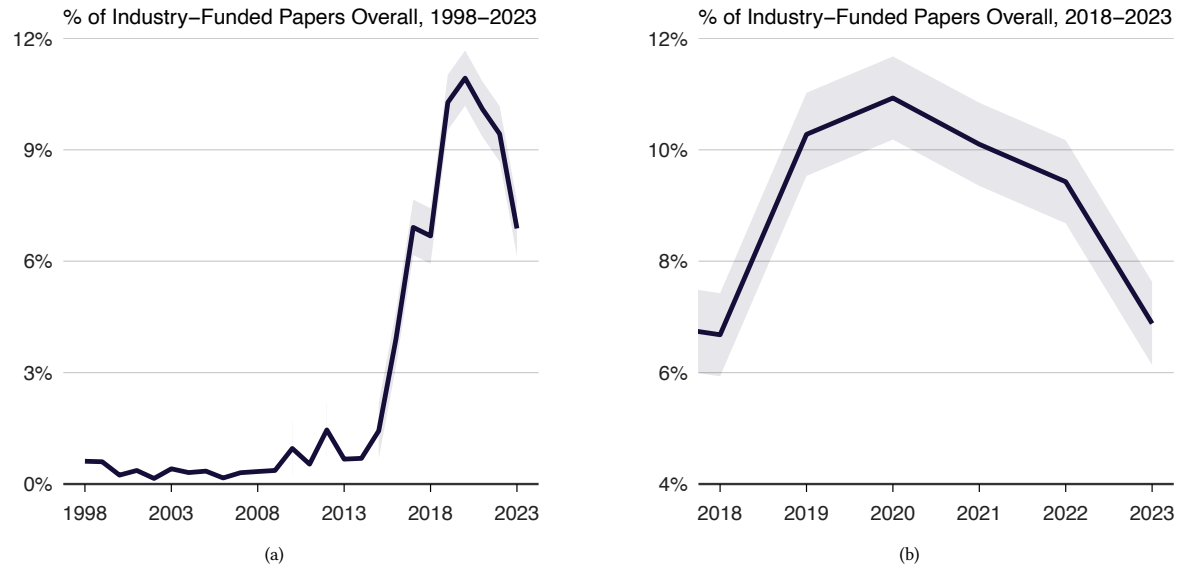


Figure 5: The *PIFP* (a) overall from 1998 to 2023 (b) and from 2018 to 2023.

biases in citation practices that are influenced by category volume. Precisely, the CPR from AI to a funding type f is defined as follows:

$$CPR_{AI}(f) = \frac{C(f)}{E(f)} \tag{3}$$

$$\text{where } C(f) = \sum_{\forall f_i \in F} C^{f_i \to f}, \tag{4}$$

$$\text{and } E(f) = \left(\sum_{\forall f_i \in F} \sum_{\forall f_j \in F} C^{f_i \to f_j} \right) \cdot \frac{N_f}{N} \tag{5}$$

where, $C(f)$ is the number of citations from all funding types to funding type f, $E(f)$ is the expected number of citations to f proportional to its share of total papers, $C^{f_i \to f_j}$ is the number of citations from funding type f_i to funding type f_j, F is the set of all subfields, and N is number of papers across all funding types. A $CPR > 1$ shows that the funding type f is more often cited by AI than expected based on its share of available papers, suggesting a positive citation preference, while a $CPR < 1$ shows that the funding type f is less often cited by AI than expected, implying a negative citation preference. A $CPR = 1$ shows that citations are proportional to availability, indicating no citation preference.

Results. Figure 6 shows the CPR of AI to different funding types over time. The CPR plot of Figure 6 reveals a consistent upward trend in citation preference towards funded papers (i.e., industry and non-industry) since 2019. By 2021, the AI community started citing more industry-funded papers than expected by the number of papers, showing an increasing reliance on industry-funded research. Despite this growing trend, non-funded research continued to be cited more frequently than industry-funded work until 2023. However, non-funded research's CPR has been steadily declining since 2019, demonstrating a reduced emphasis on non-funded research. A marked shift occurred in 2023, when citations to industry-funded papers surpassed those to non-funded papers, showing a strong preference for industry-funded research in the current AI community. Notably, while non-industry-funded research gains citations, its CPR remains negative. This behaviour shows that while the AI community is engaging more with this type of funding, the engagement is gradual and not yet proportional to its publication volume, reflecting a gradual integration of non-industry-funded research into the broader AI discourse.

Discussion. Overall, my analysis shows a marked shift in citation behaviour within the AI research community over the past five years.

The increasing CPR for both industry-funded and non-industry-funded papers demonstrates that funding has become essential to achieve academic visibility in AI research. The marked increase in the importance of industry-funded research is due to industry's contribution/development of tools and resources, such as frameworks, datasets, and models, which become foundational in AI research and development [Ahmed and Wahed 2020; Ahmed et al. 2023; Verdegem 2024]. The publications of these tools are frequently cited when used in academic and industry research alike. One example is the paper introducing PyTorch, a widely used machine learning library created by Meta [Paszke et al. 2019], which has accumulated ~49.5K[7] citations in five years. Thus, the disproportionate citation of industry-funded papers is less about sheer volume and more about citational impact, accessibility, and integration.

In response to the growing importance of AI, government agencies and non-profit organizations, such as the National Science Foundation and the European Research Council, have increased their funding in AI [André 2024]. This shift may have increased both the quantity and quality of non-industry-funded publications,

[7]Last updated 12/15/2024

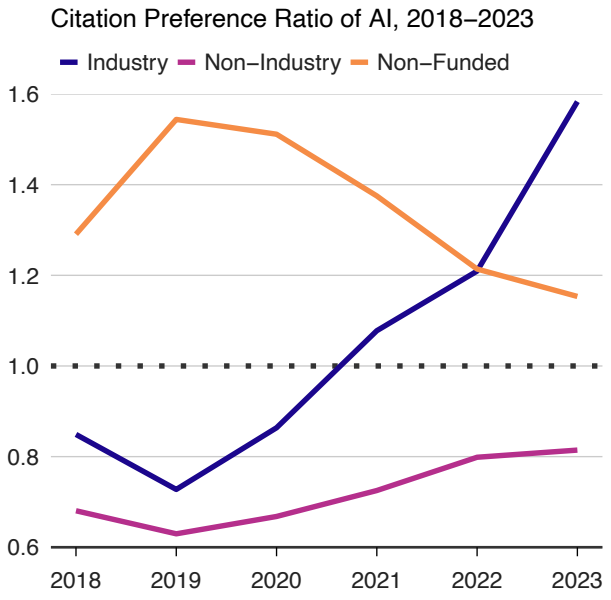


Figure 6: The Citation Preference Ratio (CPR) of AI towards industry-funded, non-industry-funded, and non-funded papers.

leading to rising citation counts. Thus, research without external funding lags behind in citation metrics, possibly due to its focus on niche topics that the broader AI community may consider less relevant.

***Q3.* To what extent do industry-funded papers cite other industry-funded papers as opposed to non-industry-funded and non-funded papers?**

Ans. Prior analysis (see **Q2.**) highlights that industry-funded research increasingly gets cited by the AI research community, but how does industry-funded research cite other types of funding? Does industry-funded research form a self-reinforcing cycle in which industry-funded research primarily cites other industry-funded work, potentially creating a shared narrative, i.e., echo chambers [Cinelli et al. 2021]? To examine this question, I calculate the difference in industry-funded outgoing citation percentage to a funding type f versus the average outgoing citations from various funding types to f. I rely on the *Outgoing Relative Citational Prominence* (*ORCP*) metric by Wahle et al. [2023b] with one key modification: instead of examining research fields, I focus on funding types. If industry-funded research (*IF*) has an ORCP greater than 0 for f, then *IF* cites f more often than other funding types cite f on average.

$$ORCP_{IF}(f) = X(f) - Y(f) \tag{6}$$

$$\text{where } X(f) = \frac{C^{IF \to f}}{\sum_{\forall f_j \in F} C^{IF \to f_j}}, \tag{7}$$

$$\text{and } Y(f) = \frac{1}{N} \sum_{i=1}^{N} \frac{C^{f_i \to f}}{\sum_{\forall f_j \in F} C^{f_i \to f_j}} \tag{8}$$

where F is the set of all funding types, N is the number of all funding types, i.e. 3, and $C^{f_i \to f_j}$ represents the number of citations from funding type f_i to funding type f_j.

Results. Figure 7 shows the ORCP scores of industry-funded papers across funding types, with industry-funded research citing itself more than average ($ORCP = 2\%$). Notably, despite the presence of extensive non-industry-funded and non-funded research of comparable quantity, both of these funding types have an $ORCP < 0$, implying that industry-funded research cites non-industry-funded and non-funded work significantly less than how much the other funding types cite non-industry-funded and non-funded research.

Figures 17 and 18 in the Appendix shows the ORCP scores for non-industry-funded and non-funded research. Among all funding types, the highest ORCP to a funding type occurs within the same funding type, indicating that citations to papers of the same funding type are more common than cross-type citations. Non-industry-funded and non-funded research has higher ORCP scores to itself than industry-funded work has to itself. Additionally, industry-funded research has the least negative ORCP among the three funding types when cited by non-industry-funded and non-funded papers.

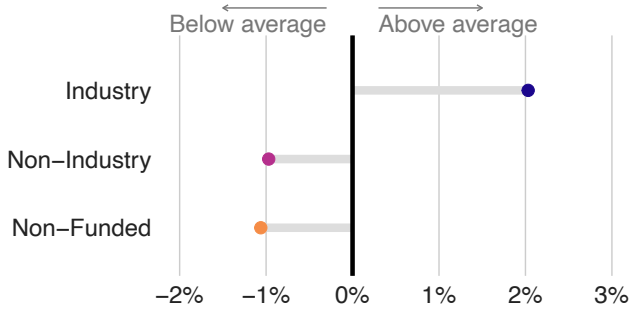


Figure 7: Industry-funded research's Outgoing Relative Citational Prominence (ORCP) scores for all funding types.

Discussion. The findings show an echo chamber effect in industry-funded research, where citations disproportionately reference similar work within the same funding ecosystem. Despite this pattern, the degree of insularity remains moderate, reflected by an ORCP of 2 %. Interestingly, all funding types demonstrate a degree of citation insularity, likely influenced by their specialized research focus. The strong thematic alignment in industry-funded research may explain the observed tendency towards self-referential citation practices.

***Q4.* How well are industry-funded papers cited? How does the citational impact vary between industry-funded, non-industry-funded, and non-funded papers?**

Ans. Despite being a minority in my dataset, industry-funded papers may have an out-sized influence on the development of AI research. To assess this effect, I examine the citational impact of industry-funded papers as a measure of influence on other researchers (i.e., non-industry and non-funded). I analyze median

citations, mean citations, and the h5-index [Hirsch 2005] across different funding types. The h5-index serves as a proxy for impact and influence, despite its known limitations in capturing all research dimensions [Bornmann and Daniel 2007; Costas and Bordons 2007].

Results. Table 4 reveals that funded papers receive more citations than non-funded papers. Non-industry-funded research has the highest h5-index (754) and the largest paper volume (29 311), demonstrating a substantial number of highly cited papers and total contributions. Conversely, industry-funded research, although smaller in quantity, achieves a disproportionately high amounts of citations, as evidenced by a substantial h5-index (503) relative to the number of papers, reflecting a focus on high-yield research outputs.

The h5-index reflects the number of papers (h) that have received at least h citations within five years [Hirsch 2005]. Industry-funded research stands out, with more than half of its publications (54 %) having high citational impact, compared to only 3 % of non-industry-funded and 2 % of non-funded papers.

Further comparisons of citation means and medians reveal variability within funding types. High means coupled with low medians show that while some papers achieve high number of citations, many do not. Industry-funded research, however, shows consistent citation patterns, suggesting a steady impact compared to other funding types, which tend to have more one-hit successes.

Table 4: The total number of AI papers published in the last five years, mean and median number of citations, as well as h5-index for different funding types are ordered by decreasing h5-index.

Funding Type	Count	Median	Mean	h5-index (↓)
Non-Industry	29 311	24	109.53	754
Industry	933	52	211.91	503
Non-Funded	22 660	5	35.70	346

Discussion. The marked amount of citations to funded papers shows a connection between research funding and high amount of citations, demonstrating the crucial role of funding in boosting research relevance and dissemination in AI research. In particular, industry-funded research shows a markedly higher citational impact relative to its publication volume compared to other funding types. This disproportionate high amount of citations may arise for various reasons, though high citation counts do not necessarily indicate perfection in every way. Nonetheless, by virtue of their visibility, highly-cited papers markedly influence research and how early researchers perceive academic writing norms [Wahle et al. 2023b].

Industry-funded papers have a self-citation bias (see **Q3.**), which can disproportionately influence early-stage researchers towards topics aligned with industry priorities. These topics are computationally intensive Maslej et al. [2023], creating incentives for researchers to pursue industry partnerships in order to gain access to exclusive resources and amplify the visibility and impact of their academic work. However, this dynamic risks shifting AI research toward profit-oriented topics aligned with industry interests, while marginalizing public needs. To mitigate this imbalance, public institutions must proactively provide researchers with the resources necessary to remain independent of industry influence. Establishing a comprehensive public research infrastructure - such as a public research cloud, public data sets, salaries, and research funding - empowers researchers to pursue work that aligns with societal needs. As Ahmed et al. [2023] argue, such infrastructure is vital not only for supporting independent research, but also for maintaining the capacity to audit industry output and ensure that AI advances serve the public interest.

Q5. **Which type of funding is most influenced by industry-funded research? How has this influence changed over the years?**

Ans. To determine the funding types most affected by industry-funded research, I analyse the citation sources to industry-funded research by funding type. Thus, I calculate the average percentage of industry-funded references per paper, i.e., the mean ratio of citations from papers with a given funding type to industry-funded papers, relative to the total citations of papers with that funding type. This approach provides a clear measure of the extent to which different funding types rely on or interact with industry-funded research over time.

Results. Figure 8 shows the proportion of outgoing citations to industry-funded papers per paper and funding type over time. It also shows the macro average of this proportion across all funding types. Observe that since 2018, the share of citations referencing industry-funded research has increased markedly across all funding types. Non-industry-funded papers showed a particularly strong growth, with 37 % increase in citations to industry-funded work per paper between 2018 and 2023 after a lower percentage start. This growth surpasses the growth of industry-funded papers (35 %) and non-funded papers (34 %). Despite this rise in cross-funding-type engagement, industry-funded papers maintained the highest proportion of outgoing citations per paper to other industry-funded research, underlining the self-referential trend of industry-funded research. However, by 2023, all funding types experienced a slight decline in the percentage of outgoing citations per paper to industry-funded papers.

Discussion. The rising proportion of outgoing citations to industry-funded papers highlights a marked increase in engagement with industry-funded research across funding types. Non-industry-funded researchers, in particular, shows growing interest in industry-driven topics and methodologies. It is still unclear why non-industry-funded research experienced that strong increase in engagement with industry-funded work. One possible reason for this engagement is the collaboration between industry and non-industry entities, with industry often providing cutting-edge resources and academia providing a platform to identify and recruit talented researchers. This perspective is supported by Klinger et al. [2022], who highlight significant industry-academia collaborations in AI research, cautioning that such partnerships may narrow thematic diversity in favor of industry-preferred topics.

The decline in outgoing citations to industry-funded papers by 2023, indicates a shift in research priorities. One can argue that the increasing interest of policymakers in AI, as noted by Maslej et al. [2023], has shifted the AI communities attention to more

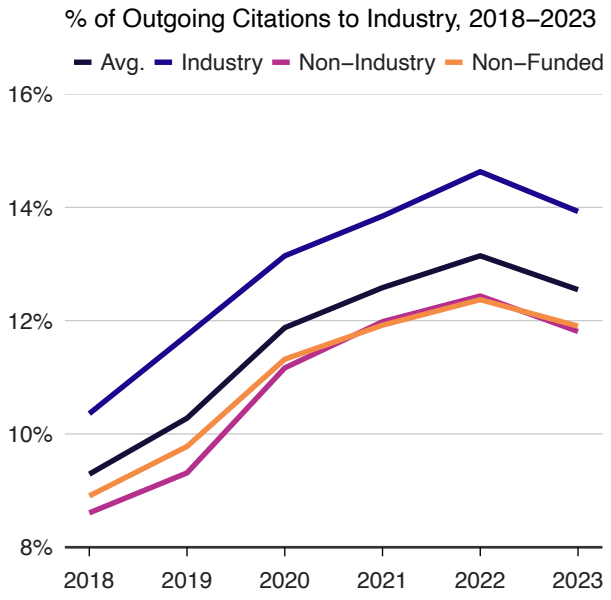


Figure 8: The average proportion of industry-funded references across various funding types for each paper. The macro-average shows the mean percentage of industry-funded references per paper over all funding types.

interdisciplinary topics that are less narrow to industry-centric applications, such as AI ethics and governance.

***Q6.* Which fields do industry-funded papers cite? How diverse are the outgoing citations in these papers, and do cited fields vary by funding type?**

Ans. To determine whether non-industry-funded and non-funded research is concentrated around industry-favoured topics, I analyze outgoing citations by funding type. Specifically, I calculate each funding type's share of citations directed to various fields, defined as the percentage of citations to a given field from a given funding type over all citations from a given funding type to any field. For papers associated with multiple fields, each field receives a citation. I use the fields pre-classified by Scopus, which classifies the fields based on the aims and scope of the title, and on the content it publishes. Serial titles are classified by Scopus in-house experts using the All Science Journal Classification scheme [Scopus 2024].

Results. Figure 9 shows the distribution of citations to the top ten cited fields for industry-funded (a), non-industry-funded (b), and non-funded papers (c). Across all funding types, eight of the ten most-cited fields belong to computer science, highlighting a strong focus on this field and a low outgoing citation field diversity. The top ten fields account for over 70 % of outgoing citations in each funding type, demonstrating a concentrated interest in these fields. Notably, industry-funded research shows the highest concentration, with 77 % of citations directed to these top ten fields, compared to 75 % for non-industry-funded papers and 72 % for non-funded papers.

The top four fields cited remain consistent across funding types, constituting more than 50 % of citations within the top ten fields, indicating a common primary interest across funding types. However, some variation exists: industry-funded papers show an increased interest in linguistics, while non-industry-funded and non-funded papers emphasize AI more prominently. Additionally, non-industry-funded research shows a stronger orientation towards theoretical work, contrasting with the industry and non-funded papers emphasis on networks and signal processing.

Discussion. The convergence of research fields across industry-funded, non-industry-funded, and non-funded research, alongside the growing engagement with industry-funded work (see **Q5.**), demonstrates the market influence of industry funding on the broader research landscape.

Building on the findings of Klinger et al. [2022], I show that industry-funded research exhibits lower thematic diversity compared to non-industry-funded and non-funded research, demonstrated by the high citation density in the top ten most-cited fields. Furthermore, my results reflect a concentration of industry-funded work in fields that are data-hungry and computationally intensive, such as computer vision and information retrieval (information systems). In contrast, industry-funded research is less focused on symbolic techniques and other theoretical aspects discussed in AI.

On the other hand, non-industry-funded and non-funded research shows a relatively strong focus on AI (concentrating on symbolic techniques) and theoretical fields such as mathematics. However, this focus does not necessarily imply that these funding types neglect data-intensive or computationally demanding fields.

***Q7.* On average, how far back in time do we go to cite AI papers? As in, what is the average age of cited papers? How does it differ across different funding types?**

Ans. To investigate the temporal patterns in scholarly citations, I adopt the methodology outlined by Bollmann and Elliott [2020]; Singh et al. [2023]; Wahle et al. [2024]. For each paper within a specific funding type, I analyse the citations to other papers and calculate how far back in time the *cited* papers were published. When a paper x cites a paper y_i, then the age of the citation (AoC) is the difference between the year of publication (YoP) of x and y_i:

$$AoC(x, y_i) = YoP(x) - YoP(y_i) \tag{9}$$

I calculate the AoC for each of the citations of a paper and average them:

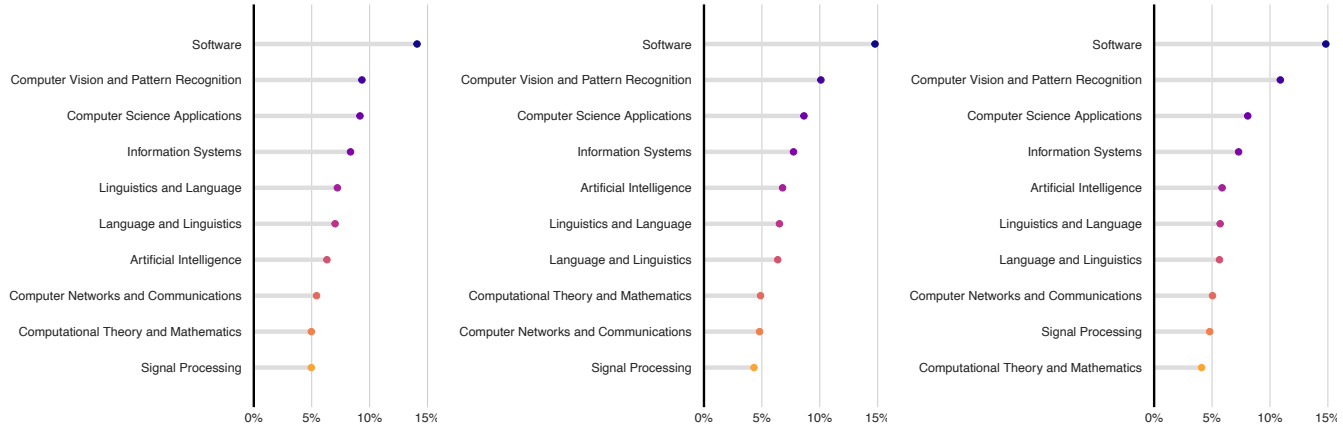
$$mAoC(x, y_i) = \frac{1}{N} \sum_{i}^{N} AoC(x, y_i) \tag{10}$$

where N denotes the total number of references in paper x.

For example, if a paper x from 2023 cites two papers, one from 2010 and one from 2020, the $mAoC$ of paper x is 8 years.

Table 5: The *mAoC* and confidence intervals for different funding types are ordered by increasing *mAoC*.

Funding Type	*mAoC* ± 95% Conf. (↑)
Industry	4.79 ± 0.02
Non-Industry	4.92 ± 0.01
Non-Funded	5.03 ± 0.03



(a) Industry (b) Non-Industry (c) Non-Funded

Figure 9: Percentage of outgoing citations from industry-funded papers (a), non-industry-funded papers (b), and non-funded papers (c) to top ten cited fields.

Results. Figure 10 shows the distribution of AoCs for all papers of each funding type and overall across the years after the publication of the *cited* paper. For example, the y-axis point for year 0 represents the average percentage of citations papers received in the same year they were published. The y-axis point for year 1 reflects the average percentage of citations received in the year following publication.

Observe that most citations occur for papers published two years prior ($AoC = 2$). For all funding types, the citation patterns show a similar trend: a sharp increase from year 0 to the peak at $AoC = 2$, followed by a decline in the years after the peak is reached. This rapid decline from the peak has a half life of about 2 years. Notably, non-funded papers have a lower peak value but maintain higher citation rates in the years after the peak compared to industry-funded and non-industry-funded papers. Additionally, industry-funded papers have the highest percentage of citations in the publication year (age 0), while non-funded papers have the lowest.

Table 5 shows the mean *mAoC* for papers published between 2018 and 2023, grouped by funding type. Observe how industry-funded papers has the lowest mean *mAoC* of 4.79, followed closely by non-industry-funded papers with a mean *mAoC* of 4.92, and non-funded papers at 5.03.

Discussion. Overall, the results show that papers typically receive the highest number of citations two years after publication, and their chances of citation fall fast after that. Non-funded papers receive fewer citations at their peak compared to industry-funded and non-industry-funded research, but their decline in citations is more gradual. This dynamic, coupled with the lower mean *mAoC* values for industry-funded and non-industry-funded papers, suggests that non-funded research is less fast paced than other types of funding, likely reflecting different research priorities.

Non-funded research may be more focussed on foundational or theoretical contributions that continue to attract citations over time. In contrast, industry-funded and non-industry-funded research tends to focus on more recent and rapidly evolving innovations,

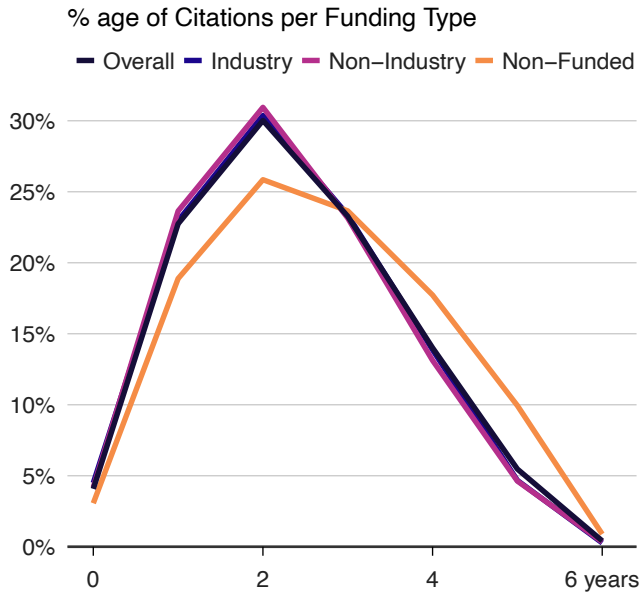


Figure 10: Distribution of *AoC* for papers in AI (overall and by funding type).

reflecting the fast-paced nature of technological advancements and their disruptional potential. This divergence highlights distinct time dynamics and the varying impact of funding types on the longevity of scientific contributions.

However, the extent to which these trends have persisted historically remains unclear. Understanding whether industry-funded research has always had such a low citation age, and how these citation trends have evolved over time, requires further analysis.

***Q8.* What is the distribution of *mAoC* in industry-funded papers? How does this distribution vary across years?**

Ans. To answer this question, I calculate the *mAoC* for each industry-funded paper, as well as the median and mean *mAoC* for all industry-funded papers over time. If a paper x was published in year t, then $mAoC(x)$ contributes to the distribution for year t.

Results. Figure 11 shows the violin plots for distributions of *mAoC* in industry-funded papers across various years. Each plot highlights the median (marked with a white diamond within the grey rectangle), representing the recency of citations for that year. Over time, the median *mAoC* for industry-funded papers shows a consistent decline, indicating an increasing focus on citing relatively recent work. The shrinking size of the second and third quartiles (halves of the grey rectangle), indicates that citations are increasingly concentrated around the median, reflecting a narrowing range in citation age.

From 2018 to 2023, the mean *mAoC* closely follows the median trend but, remains consistently higher, revealing a right skew in the data. This skew is due to a number of papers citing much older papers, which markedly affect the mean. Additionally, the decreasing standard deviation suggesting diminishing citation age diversity, possibly reinforcing the trend toward citing newer literature.

The violin plots visually capture this evolution. By 2023, the violin's density transforms into a spinning tractroid top[8]. This transformation reflects the increasing concentration of *mAoC* values near five years, with a high fraction of papers clustering tightly around this point.

Discussion. The results show a decline in citations to older works within industry-funded papers, accompanied by a reduction in the temporal diversity of citations. Although the exact causes of this trend remain uncertain, multiple factors contribute to the evolving citation dynamics. The substantial impact of transformers on NLP and ML, as well as academic incentives, shaped by the preferences of reviewers, institutions, and conferences, can increasingly favor more recent publications. This trend reflects evolving priorities within the academic community. The pressures of the "publish or perish" paradigm further exacerbate this trend, encouraging researchers to divide their work into smaller, publishable units. Additionally, the rise of open-access initiatives and preprint servers, which provide immediate access to research, has amplified the tendency to cite newer works.

My results add to (and are consistent with) the mean-citation age results found by Wahle et al. [2024], who analyzed the mean citation age of NLP papers from 1990 to 2023. By focusing on AI publications between 2018 and 2023, my analyses situates those results in the overall trajectory of how temporal citation patterns in AI have evolved since the impact of transformer-based models to the present period.

5 CONCLUDING REMARKS

This work examined the citational impact of Big Tech funding on AI research through a set of comprehensive analyses of citational patterns in scientific literature. To enable this analysis, I compiled a unique dataset of metadata that includes ~57.3K AI papers, their funding agencies (if any), citations to AI papers, and citations by the AI papers. I analyzed this data using various metrics such as *Citation Preference Ratio*, *Relative Citational Prominence*, and *Mean Age of Citation* to show a growing tendency within the AI research community to engage with industry-funded work. However, this trend comes at the expense of diversity, as industry-funded research disproportionately cites itself while neglecting a broader range of older, potentially foundational work.

My findings show a paradox: while the presence of industry-funded research in top AI conferences is declining, its citational influence continues to grow. Contributions from non-industry-funded and non-funded research receive little recognition in industry-funded research, despite the convergence in research fields across funding types. This dynamic reflects the growing insularity of industry-funded research. My experiments also show that the diversity of age of citations and the percentage of older papers cited by industry-funded papers have declined. This decline risks losing valuable insights and principles essential for fostering responsible and inclusive technological development.

Over the past five years, the widespread adoption of AI technologies has directly and indirectly affected billions of lives, sometimes with markedly negative consequences due to a lack of foresight in system development. It is well-documented that engaging with diverse literature, spanning multiple disciplines and time periods, is critical for creating more inclusive systems. Yet, the observed tendency of industry-funded papers to favor recent, similarly funded work over diverse and older references signals a move toward an increasingly insular research culture. If left unaddressed, this trend could impact the development of technologies that are not beneficial for all, but for those with power.

The scientific community has a unique responsibility to counteract these trends. It is a fallacy to accept (temporal) citation patterns as inevitable, or to assume that researchers lack agency in their citation choices. By reflecting on our reading habits and engaging with a wider array of literature, we can foster a more diverse and inclusive scientific discourse. Public institutions play a critical role in this effort by improving their funding policy's and providing researchers with the tools and support they need to maintain independence from industry influence and promote diverse citation practices. This collective commitment is essential to ensure that the development of AI technologies prioritizes public benefit over profit-driven objectives.

6 LIMITATIONS

6.1 Manual Analysis

The manual analysis has a few limitations. First, because examining thousands of funding agencies individually is a time-intensive process, this analysis only includes only 5% of the extracted funding agencies. Notably, this 5% covers 74% of all funding occurrences, providing robust overall representation. Second, identifying CFs relied on available online information. In cases where insufficient data prevented me from confirming an industry affiliation, I marked agencies as non-funded, potentially leading to false negatives in the dataset's metadata. Third, as this analysis was conducted solely by

[8]Form of the iconic spinning top in the movie Inception.

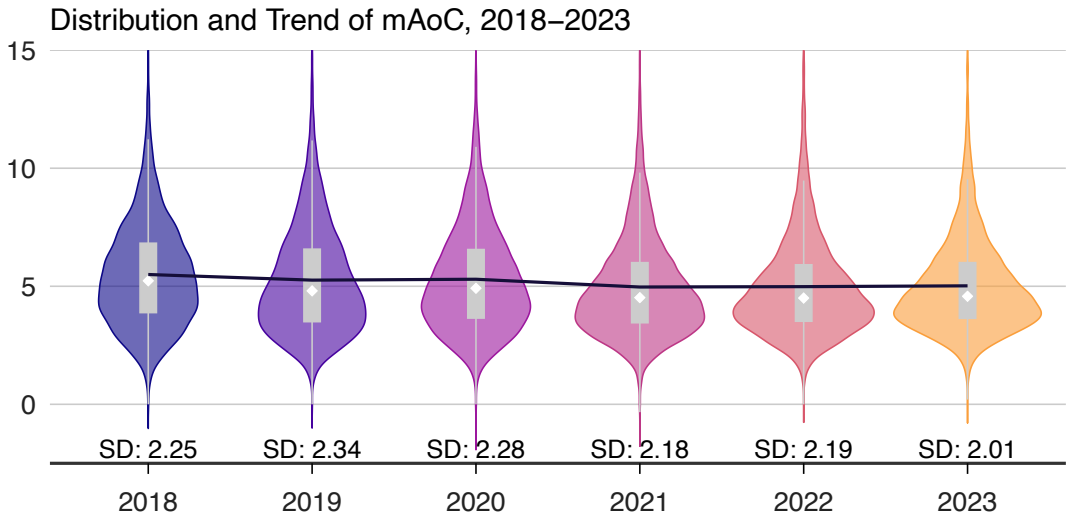


Figure 11: Distribution of *mAoC* for industry-funded papers between 2018 and 2023. The standard deviation for each year is displayed below the respective violin plot. The median (white diamond) and the mean *mAoC* (dark line) are shown for each year.

one person, interpretation variability may affect data consistency and quality.

6.2 Automatic Analysis

Identifying CFs through automated, fuzzy text matching with standardized company names can lead to false positives, where unrelated agencies are incorrectly matched. To mitigate this risk of false positives, a high similarity threshold (90%) was set, although this conservative threshold could reduce the number of matches and potentially miss some CFs. However, most funding agency names included the standardized identifier (e.g., Google, Google DeepMind, Google Cloud), minimizing the risk of missing relevant CFs.

6.3 General Limitations

Beyond technical limitations, this thesis faces broader constraints. Industry involvement in AI research today extends beyond direct financial contributions to include access to models, datasets, computational resources, and specialized expertise [Ahmed et al. 2023; Montes and Goertzel 2019; Riedl 2020; Verdegem 2024]. This analysis captures industry influence solely through funding data from Scopus, which is based on paper acknowledgments [Liu 2020]. The extent of funding information in these acknowledgments varies depending on the details provided in the publications, reflecting disciplinary and regional disparities in funding reporting practices [Pranckutė 2021]. Notably, studies by Liu [2020] and Pranckutė [2021] identify errors in Scopus funding acknowledgment text and funding agency fields. Therefore, the consistency and quality of the identified industry presence should be taken with a grain of salt.

This research focuses on publications from prominent AI-related conferences, rather than all AI-related academic publications. Although leading conferences shape the academic research agenda [Freyne et al. 2010], this selection excludes vibrant, often non-English AI communities and venues, limiting the generalizability of my findings to the global AI research landscape. Future studies are needed to explore industry influence across diverse subcommunities and venues worldwide.

Furthermore, this analysis quantifies influence primarily through citations, a method with inherent limitations. Citation counts alone lack nuance, as not all citations reflect the same level of influence [Valenzuela et al. 2015; Zhu et al. 2015]. Additionally, citation patterns are affected by biases [Ioannidis et al. 2019; Mohammad 2020b; Nielsen and Andersen 2021]. This work also examines citation practices on a large scale, focusing on quantitative trends. Qualitative aspects may reveal the reasons behind why industry-funded research receives more engagement within the AI community, shows growing insularity, and cites recent over older literature. Several factors may contribute to this, such as the volume of recent publications, the applicability of industry-funded research, and the technical relevance of industry-funded work.

Another aspect that my analysis did not address is the allocation of financial resources by industry across AI subfields and conferences. An analysis of the cash flows from industry to AI research and their impact over time could reveal whether financial resources markedly drive influential AI research or whether other resources are key. Exploring the cash flows, their impact, and their presence over time is an area I leave to future work.

7 ETHICAL CONSIDERATIONS

This thesis conducts an analysis of scientific literature at an aggregate level, using data from the Scopus database. The database provides metadata such as titles, authors, funding agencies, and publication years, all of which are used without infringing on copyrighted content. All of the analyses in this thesis are at aggregate-level, and not about individual papers or authors.

A critical aspect of this thesis is its reliance on citation counts as a proxy to characterize funding types. While citations serve as a convenient metric, this approach raises concerns about potential misinterpretation or misuse of my findings. For example, the observed high number of outgoing citations to industry-funded research should not be used as a rationale for diminishing research funded by non-industry sources or conducted without external funding. To address the risks of oversimplified interpretations, a more comprehensive evaluation framework may be beneficial. Such a framework would integrate multiple dimensions, including relevance, popularity, resource availability, impact, geographic context, and temporal trends, thus mitigating the problems of shallow analysis.

ACKNOWLEDGMENTS

The thesis received no external funding. Many thanks to Dr. Terry Lima Ruas and Jan Philip Wahle for their valuable discussions, feedback, ideas, and help in developing and guiding this thesis.

REFERENCES

Mohamed Abdalla and Moustafa Abdalla. 2021. The Grey Hoodie Project: Big Tobacco, Big Tech, and the Threat on Academic Integrity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 287–297. https://doi.org/10.1145/3461702.3462563

Mohamed Abdalla, Jan Philip Wahle, Terry Ruas, Aurélie Névéol, Fanny Ducel, Saif Mohammad, and Karen Fort. 2023b. The Elephant in the Room: Analyzing the Presence of Big Tech in Natural Language Processing Research. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 13141–13160. https://doi.org/10.18653/v1/2023.acl-long.734

Salwa Abdalla, Moustafa Abdalla, Mohamed Saad, David Jones, Scott Podolsky, and Mohamed Abdalla. 2023a. Ethnicity and gender trends of UK authors in The British Medical Journal and the Lancet over the past two decades: a comprehensive longitudinal analysis. *EClinicalMedicine* 64 (2023).

Nur Ahmed and Muntasir Wahed. 2020. The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research. arXiv:2010.15581 [cs.CY]

Nur Ahmed, Muntasir Wahed, and Neil C Thompson. 2023. The growing influence of industry in AI research. *Science* 379, 6635 (2023), 884–886.

MADIEGA Tambiama André. 2024. AI investment: EU and global indicators. (2024).

Ian Ayres and Fredrick E Vars. 2000. Determinants of citations to articles in elite law reviews. *The Journal of Legal Studies* 29, S1 (2000), 427–450.

Marcel Bollmann and Desmond Elliott. 2020. On Forgetting to Cite Older Papers: An Analysis of the ACL Anthology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 7819–7827. https://doi.org/10.18653/v1/2020.acl-main.699

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).

Lutz Bornmann and Hans-Dieter Daniel. 2007. What do we know about the h index? *Journal of the American Society for Information Science and technology* 58, 9 (2007), 1381–1385.

Gualberto Buela-Casal and Izabela Zych. 2010. Analysis of the relationship between the number of citations and the quality evaluated by experts in psychology journals. *Psicothema* (2010), 270–276.

Paula Chatterjee and Rachel M Werner. 2021. Gender disparity in citations in high-impact journal articles. *JAMA Network Open* 4, 7 (2021), e2114509–e2114509.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, 9 (2021), e2023301118.

Iain M Cockburn, Rebecca Henderson, and Scott Stern. 2018. *The Impact of Artificial Intelligence on Innovation.* Working Paper 24449. National Bureau of Economic Research. https://doi.org/10.3386/w24449

François Collet, Duncan A Robertson, and Daniela Lup. 2014. When does brokerage matter? Citation impact of research teams in an emerging academic field. *Strategic Organization* 12, 3 (2014), 157–179.

Rodrigo Costas and María Bordons. 2007. The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of informetrics* 1, 3 (2007), 193–203.

Rodrigo Costas, Maria Bordons, Thed N Van Leeuwen, and Anthony FJ Van Raan. 2009. Scaling rules in the science system: Influence of field-specific citation characteristics on the impact of individual researchers. *Journal of the American Society for Information Science and Technology* 60, 4 (2009), 740–753.

Derek John de Solla Price. 1962. *Science since babylon.* Yale University Press New Haven, CT.

Sergio Della Sala and Joanna Brooks. 2008. Multi-authors' self-citation: A further impact factor bias? *Cortex; a journal devoted to the study of the nervous system and behavior* 44, 9 (2008), 1139–1145.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] https://arxiv.org/abs/1810.04805

Roel Dobbe and Meredith Whittaker. 2019. AI and climate change: how they're connected, and what we can do about it. *AI Now Institute* 17 (2019).

Matthew E Falagas, Angeliki Zarkali, Drosos E Karageorgopoulos, Vangelis Bardakas, and Michael N Mavros. 2013. The impact of article length on the number of future citations: a bibliometric analysis of general medicine journals. *PloS one* 8, 2 (2013), e49476.

Jill Freyne, Lorcan Coyle, Barry Smyth, and Padraig Cunningham. 2010. Relative status of journal and conference publications in computer science. *Commun. ACM* 53, 11 (nov 2010), 124–132. https://doi.org/10.1145/1839676.1839701

Bela Gipp and Norman Meuschke. 2011. Citation pattern matching algorithms for citation-based plagiarism detection: greedy citation tiling, citation chunking and longest common citation sequence. In *Proceedings of the 11th ACM symposium on Document engineering*. 249–258.

Jorge E Hirsch. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences* 102, 46 (2005), 16569–16572.

John PA Ioannidis, Jeroen Baas, Richard Klavans, and Kevin W Boyack. 2019. A standardized citation metrics author database annotated for scientific field. *PLoS biology* 17, 8 (2019), e3000384.

Joel Klinger, Juan Mateos-Garcia, and Konstantinos Stathoulopoulos. 2022. A narrowing of AI research? arXiv:2009.10385 [cs.CY] https://arxiv.org/abs/2009.10385

Mario Krenn, Lorenzo Buffoni, Bruno Coutinho, Sagi Eppel, Jacob Gates Foster, Andrew Gritsevskiy, Harlin Lee, Yichao Lu, João P. Moutinho, Nima Sanjabi, Rishi Sonthalia, Ngoc Mai Tran, Francisco Valente, Yangxinyu Xie, Rose Yu, and Michael Kopp. 2023. Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network. *Nature Machine Intelligence* 5, 11 (Oct. 2023), 1326–1335. https://doi.org/10.1038/s42256-023-00735-0

Caitlin Kuhlman, Latifa Jackson, and Rumi Chunara. 2020. No computation without representation: Avoiding data and algorithm biases through diversity. arXiv:2002.11836 [cs.CY]

Rodrigo Pessoa Cavalcanti Lira, Rafael Marsicano Cezar Vieira, Fauze Abdulmassih Gonçalves, Maria Carolina Alves Ferreira, Diana Maziero, Thais Helena Moreira Passos, and Carlos Eduardo Leite Arieta. 2013. Influence of English language in the number of citations of articles published in Brazilian journals of Ophthalmology. *Arquivos Brasileiros de Oftalmologia* 76 (2013), 26–28.

Weishu Liu. 2020. Accuracy of funding information in Scopus: A comparative case study. *Scientometrics* 124, 1 (2020), 803–811.

Anaïs Llorens, Athina Tzovara, Ludovic Bellier, Ilina Bhaya-Grossman, Aurélie Bidet-Caulet, William K Chang, Zachariah R Cross, Rosa Dominguez-Faus, Adeen Flinker, Yvonne Fonken, et al. 2021. Gender bias in academia: A lifetime problem that needs solutions. *Neuron* 109, 13 (2021), 2047–2074.

Dieuwertje Luitse and Wiebke Denkena. 2021. The great Transformer: Examining the role of large language models in the political economy of AI. *Big Data & Society* 8, 2 (2021), 20539517211047734. https://doi.org/10.1177/20539517211047734 arXiv:https://doi.org/10.1177/20539517211047734

Joseph Mariani, Gil Francopoulo, and Patrick Paroubek. 2019. The nlp4nlp corpus (i): 50 years of publication, collaboration and citation in speech and language processing. *Frontiers in Research Metrics and Analytics* 3 (2019), 36.

Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault. 2023. Artificial Intelligence Index Report 2023. arXiv:2310.03715 [cs.AI] https://arxiv.org/abs/2310.03715

Saif Mohammad. 2020a. NLP scholar: A dataset for examining the state of NLP research. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 868–877.

Saif M. Mohammad. 2020b. Examining Citations of Natural Language Processing Literature. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5199–5209. https://doi.org/10.18653/v1/2020.acl-main.464

Saif M Mohammad. 2020c. Gender gap in natural language processing research: Disparities in authorship and citations. *arXiv preprint arXiv:2005.00962* (2020).

Gabriel Axel Montes and Ben Goertzel. 2019. Distributed, decentralized, and democratized artificial intelligence. *Technological Forecasting and Social Change* 141 (2019), 354–358.

J.P. Morgan. 2023. The future of Big Tech | J.P. Morgan Research. https://www.jpmorgan.com/insights/global-research/technology/future-of-big-tech
Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. arXiv:2004.09456 [cs.CL]
Hoa Nguyen and Steffen Eger. 2024. Is there really a Citation Age Bias in NLP? arXiv:2401.03545 [cs.DL] https://arxiv.org/abs/2401.03545
Mathias Wullum Nielsen and Jens Peter Andersen. 2021. Global citation inequality is on the rise. *Proceedings of the National Academy of Sciences* 118, 7 (2021), e2012208118.
Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv:1912.01703 [cs.LG] https://arxiv.org/abs/1912.01703
Raminta Pranckutė. 2021. Web of Science (WoS) and Scopus: The Titans of Bibliographic Information in Today's Academic World. *Publications* 9, 1 (2021). https://doi.org/10.3390/publications9010012
Mark Riedl. 2020. AI democratization in the era of GPT-3. *The Gradient* 25 (2020).
Ludovic Righetti, Raj Madhavan, and Raja Chatila. 2019. Unintended consequences of biased robotic and artificial intelligence systems [ethical, legal, and societal issues]. *IEEE Robotics & Automation Magazine* 26, 3 (2019), 11–13.
Mukund Rungta, Janvijay Singh, Saif M Mohammad, and Diyi Yang. 2022. Geographic citation gaps in NLP research. *arXiv preprint arXiv:2210.14424* (2022).
Scopus. 2024. What are Scopus subject area categories and ASJC codes? https://service.elsevier.com/app/answers/detail/a_id/12007/supporthub/scopus/
Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. 2022. Compute trends across three eras of machine learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.
Sei-Ching Joanna Sin. 2011. International coauthorship and citation impact: A bibliometric study of six LIS journals, 1980–2008. *Journal of the American Society for Information Science and Technology* 62, 9 (2011), 1770–1783.
Janvijay Singh, Mukund Rungta, Diyi Yang, and Saif M Mohammad. 2023. Forgotten knowledge: Examining the citational amnesia in NLP. *arXiv preprint arXiv:2305.18554* (2023).
Pippa Smart. 2022. The evolution, benefits, and challenges of preprints and their interaction with journals. *Science Editing* 9, 1 (2022), 79–84.
Mike Thelwall, Subreena Simrick, Ian Viney, and Peter Van den Besselaar. 2023. What is research funding, how does it influence research, and how is it recorded? Key dimensions of variation. *Scientometrics* 128, 11 (2023), 6085–6106.
Patrice Trouiller, Piero Olliaro, Els Torreele, James Orbinski, Richard Laing, and Nathan Ford. 2002. Drug development for neglected diseases: a deficient market and a public-health policy failure. *The Lancet* 359, 9324 (2002), 2188–2194.
Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying meaningful citations. In *Workshops at the twenty-ninth AAAI conference on artificial intelligence*.
A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).
Pieter Verdegem. 2024. Dismantling AI capitalism: the commons as an alternative to the power concentration of Big Tech. *AI & society* 39, 2 (2024), 727–737.
Alex Verstak, Anurag Acharya, Helder Suzuki, Sean Henderson, Mikhail Iakhiaev, Cliff Chiung Yu Lin, and Namit Shetty. 2014. On the Shoulders of Giants: The Growing Impact of Older Articles. arXiv:1411.0275 [cs.DL] https://arxiv.org/abs/1411.0275
Jan Philip Wahle, Bela Gipp, and Terry Ruas. 2023a. Paraphrase Types for Generation and Detection. *arXiv preprint arXiv:2310.14863* (2023).
Jan Philip Wahle, Terry Ruas, Mohamed Abdalla, Bela Gipp, and Saif M Mohammad. 2023b. We are who we cite: Bridges of influence between natural language processing and other academic fields. *arXiv preprint arXiv:2310.14870* (2023).
Jan Philip Wahle, Terry Ruas, Mohamed Abdalla, Bela Gipp, and Saif M Mohammad. 2024. Citation Amnesia: NLP and Other Academic Fields Are in a Citation Age Recession. *arXiv preprint arXiv:2402.12046* (2024).
Jan Philip Wahle, Terry Ruas, Tomáš Foltýnek, Norman Meuschke, and Bela Gipp. 2022a. Identifying machine-paraphrased plagiarism. In *International Conference on Information*. Springer, 393–413.
Jan Philip Wahle, Terry Ruas, Saif M Mohammad, and Bela Gipp. 2022b. D3: A massive dataset of scholarly metadata for analyzing the state of computer science research. *arXiv preprint arXiv:2204.13384* (2022).
Jan Philip Wahle, Terry Ruas, Saif M. Mohammad, Norman Meuschke, and Bela Gipp. 2023c. AI Usage Cards: Responsibly Reporting AI-generated Content. arXiv:2303.03886 [cs.CY] https://arxiv.org/abs/2303.03886
Jue Wang and Philip Shapira. 2015. Is there a relationship between research sponsorship and publication impact? An analysis of funding acknowledgments in nanotechnology papers. *PloS one* 10, 2 (2015), e0117727.
Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. Discriminating systems. *AI Now* (2019), 1–33.
Meredith Whittaker. 2021. The steep cost of capture. *Interactions* 28, 6 (2021), 50–55.
Xiaodan Zhu, Peter Turney, Daniel Lemire, and André Vellino. 2015. Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology* 66, 2 (2015), 408–427.

A APPENDIX

A.1 Details on the Extraction of Companies

I searched for company names and common aliases (e.g., Microsoft, Microsoft Azure, Microsoft Cloud Computing Research Centre) using the fuzzywuzzy python package[9] with a 90 % threshold.

Table 6: Company name standardization.

Names of funding agencies	Std. Name
Microsoft, Microsoft Azure, Microsoft Research	Microsoft
Amazon, AWS, Amazon Research	Amazon
Google, Google DeepMind, Google Cloud	Google
Nvidia, NVIDIA AI Center, NVIDIA Corp	Nvidia

A.2 Supplemental Experimental Results

In addition to the primary results presented in this thesis, I describe supplementary results in the form of additional statistics and plots.

A.2.1 Extended Results on Industry Presence in AI.

Figure 12 shows the FIFP from **Q1.** for the time frame 2018 to 2023. Figure 13 shows the percentage of industry-funded papers in an AI subfield out of all industry-funded papers.

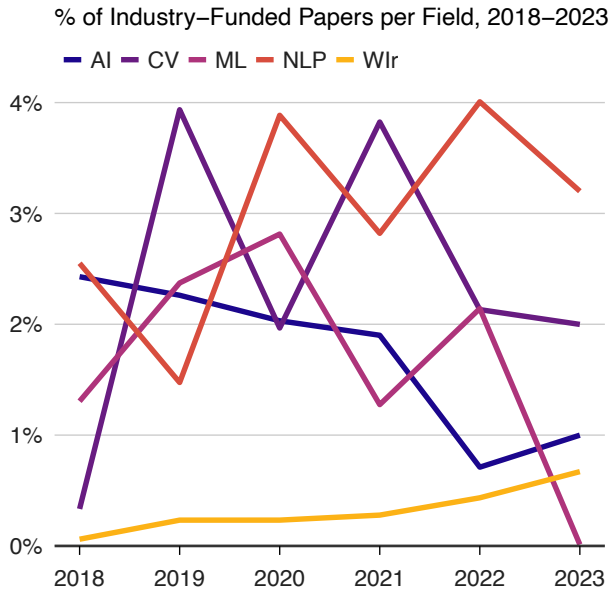


Figure 12: The *FIFP* from 2018 to 2023.

A.2.2 Extended Results on Citation Preference Ratio.

Figures 14 to 16 shows the CPR of industry-funded, non-industry-funded, and non-funded papers to different funding types over time.

A.2.3 Extended Results on Relative Citation Prominence.

Figure 17 shows the ORCP from **Q3.** for non-industry-funded papers, and Figure 18 shows the ORCP for non-funded papers.

[9] https://pypi.org/project/fuzzywuzzy/

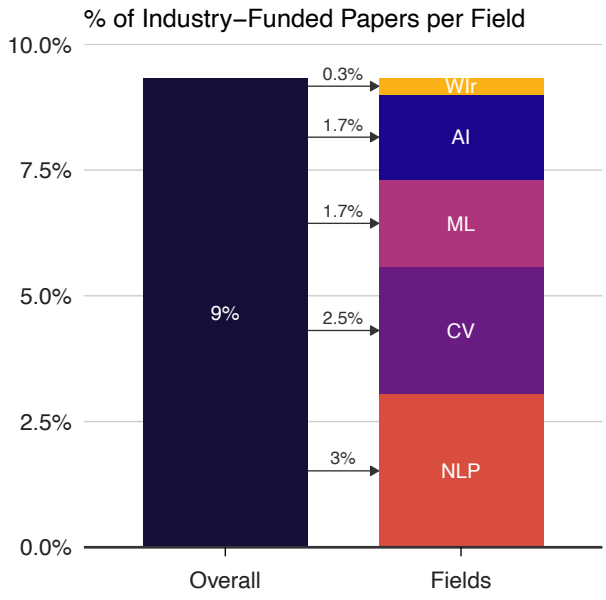


Figure 13: The percentage share of AI subfields in papers funded (2018-2023), in ascending order.

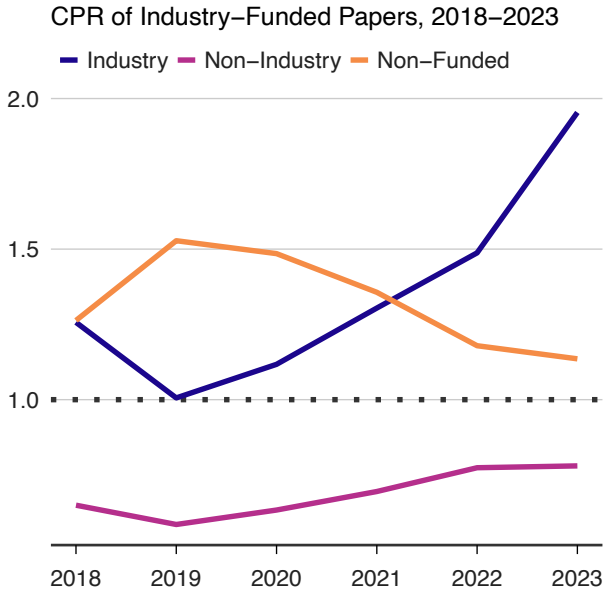


Figure 14: The Citation Preference Ratio (CPR) of industry-funded papers towards industry-funded, non-industry-funded, and non-funded papers.

A.3 AI Usage Card

I report how I used AI assistants such as ChatGPT and Claude for this work in the following standardized card according to Wahle et al. [2023c].

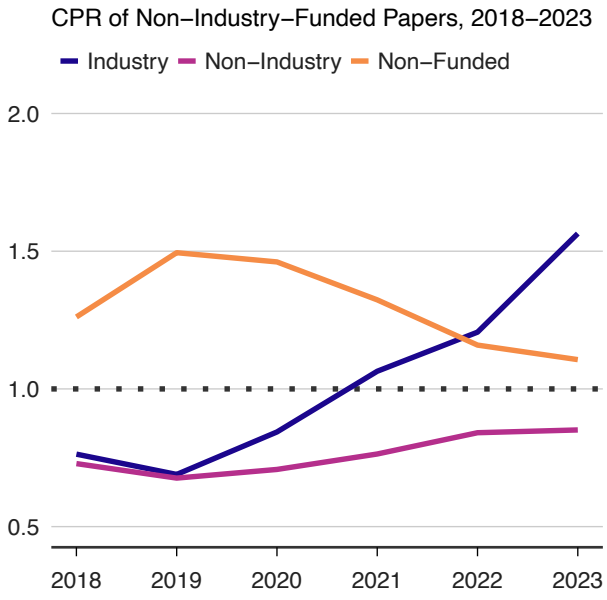


Figure 15: The Citation Preference Ratio (CPR) of non-industry-funded papers towards industry-funded, non-industry-funded, and non-funded papers.

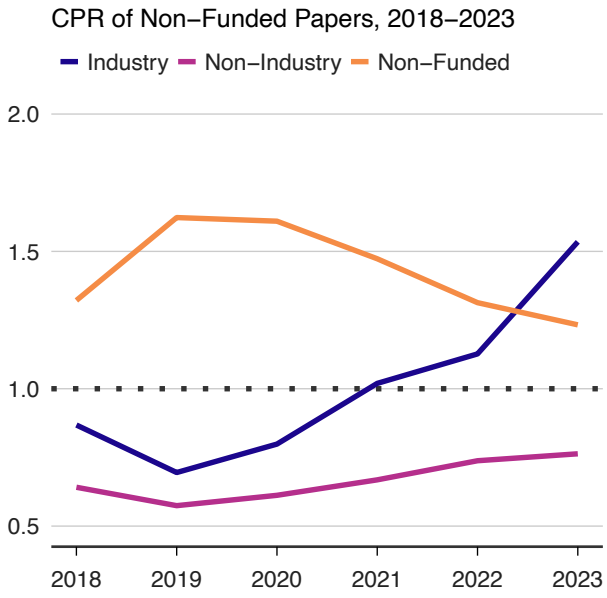


Figure 16: The Citation Preference Ratio (CPR) of non-funded papers towards industry-funded, non-industry-funded, and non-funded papers.

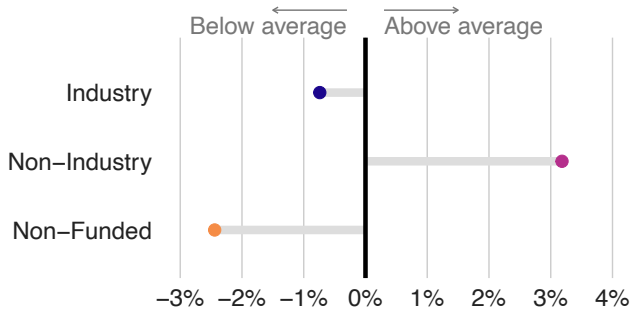


Figure 17: Non-industry-funded research's Outgoing Relative Citational Prominence (ORCP) scores for all funding types.

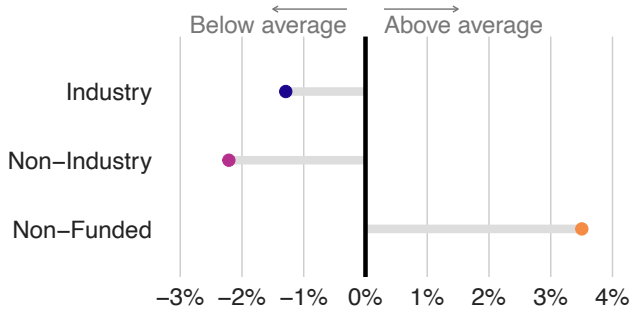


Figure 18: Non-funded research's Outgoing Relative Citational Prominence (ORCP) scores for all funding types.

AI Usage Card

CORRESPONDENCE(S)
Max Martin Gnewuch

CONTACT(S)
maxmartin.gnewuch@stud.uni-goettingen.de

AFFILIATION(S)
University of Göttingen

PROJECT NAME
What Impact Does Big Tech Funding Have on AI Research? A Scholarly Document Analysis

KEY APPLICATION(S)
Scientometrics, Citation Analysis, Artificial intelligence, Research funding, Academic research funding, Industry influence, Diversity, Fairness

MODEL(S)	DATE(S) USED	VERSION(S)
ChatGPT	2024-04-01	4o, 4o1
Claude	2024-05-01	3.5 Sonnet

IDEATION

GENERATING IDEAS, OUTLINES, AND WORKFLOWS
Not used

IMPROVING EXISTING IDEAS
Not used

FINDING GAPS OR COMPARE ASPECTS OF IDEAS
Not used

LITERATURE REVIEW

FINDING LITERATURE
Not used

FINDING EXAMPLES FROM KNOWN LITERATURE
Not used

ADDING ADDITIONAL LITERATURE FOR EXISTING STATEMENTS AND FACTS
Not used

COMPARING LITERATURE
Not used

METHODOLOGY

PROPOSING NEW SOLUTIONS TO PROBLEMS
Not used

FINDING ITERATIVE OPTIMIZATIONS
Not used

COMPARING RELATED SOLUTIONS
Not used

EXPERIMENTS

DESIGNING NEW EXPERIMENTS
Not used

EDITING EXISTING EXPERIMENTS
Not used

FINDING, COMPARING, AND AGGREGATING RESULTS
Not used

WRITING
ChatGPT Claude

GENERATING NEW TEXT BASED ON INSTRUCTIONS
Used

ASSISTING IN IMPROVING OWN CONTENT
Used

PARAPHRASING RELATED WORK
Used

PUTTING OTHER WORKS IN PERSPECTIVE
Not used

PRESENTATION

GENERATING NEW ARTIFACTS
Not used

IMPROVING THE AESTHETICS OF ARTIFACTS
Not used

FINDING RELATIONS BETWEEN OWN OR RELATED ARTIFACTS
Not used

CODING
ChatGPT Claude

GENERATING NEW CODE BASED ON DESCRIPTIONS OR EXISTING CODE
Used

REFACTORING AND OPTIMIZING EXISTING CODE
Used

COMPARING ASPECTS OF EXISTING CODE
Not used

DATA

SUGGESTING NEW SOURCES FOR DATA COLLECTION
Not used

CLEANING, NORMALIZING, OR STANDARDIZING DATA
Not used

FINDING RELATIONS BETWEEN DATA AND COLLECTION METHODS
Not used

ETHICS

WHAT ARE THE IMPLICATIONS OF USING AI FOR THIS PROJECT?
Generating code and improving the clearity of writing the paper has improved the efficacy of performing this scientific work.

WHAT STEPS ARE WE TAKING TO MITIGATE ERRORS OF AI FOR THIS PROJECT?
I manually fact-checked generated texts and inspected source code for potential generated bugs.

WHAT STEPS ARE WE TAKING TO MINIMIZE THE CHANCE OF HARM OR INAPPROPRIATE USE OF AI FOR THIS PROJECT?
I did not include text suggestions that had any chance of impacting marginalized groups.

THE CORRESPONDING AUTHORS VERIFY AND AGREE WITH THE MODIFICATIONS OR GENERATIONS OF THEIR USED AI-GENERATED CONTENT
Yes