On Tables with Numbers, with Numbers

Konstantinos Kogkalidis
Aalto University
Department of Computer Science
kokos.kogkalidis@aalto.fi

Stergios Chatzikyriakidis
University of Crete
Department of Philology
stergios.chatzikyriakidis@uoc.gr

Abstract

This paper is a critical reflection on the epistemic culture of contemporary computational linguistics, framed in the context of its growing obsession with tables with numbers. We argue against tables with numbers on the basis of their epistemic irrelevance, their environmental impact, their role in enabling and exacerbating social inequalities, and their deep ties to commercial applications and profit-driven research. We substantiate our arguments with empirical evidence drawn from a meta-analysis of computational linguistics research over the last decade.

1 Introduction

Throughout its evolution, computational linguistics has undergone multiple identity crises. In its present from, and despite its logical origins and linguistic ambitions, it is almost entirely aligned with positivist principles and ideals (Church and Liberman, 2021). The imprint of this alignment is an idealization of experimental quantification, most commonly manifesting in the form of *tables with numbers*. Tables with numbers can certainly be useful. That said, their centrality in contemporary computational linguistics research is indicative of both scientific reductionism and technological obsession. Beneath the numbers lie signs of a field in disarray: a waning reliance on theory (linguistic or otherwise), nowadays substituted by model scale; a disproportionate representation of big industry and big academia, in turn associated with a lack of transparency, accessibility and inclusion; an experimental paradigm dominated by stagnant "task-and-benchmark" practices, detached from technical rigor as well as scientific insight; and a progressive estrangement from societal, humanistic and environmental context. And while the community seems to be both alert to and uneasy with the current state of affairs (Michael et al., 2023; Gururaja et al., 2023), a holistic analysis of these issues has been long missing from the literature.

In this paper, we brave a look under the number rock. We conduct a critical assessment of the epistemic culture of computational linguistics, focusing specifically on its relation to tables with numbers. We narrow down on four axes of interest:

- The epistemological preconditions that granted tables with numbers the status of scientific currency, and the mechanisms that affect their actual value (§2).
- Their environmental footprint and the normative discourse around it (§3).
- Their cause-and-effect relation to the perpetuation and exacerbation of inequality and harmful power structures (§4).
- Their intrinsic ties with corporate interest, profit, and the accumulation of technoscientific capital (§5).

2 The Multiple Facets of Number

The field's dominant scientific approach embodies a wildly exaggerated version of positivism. This is evident both in the themes prevalent in the mainstream discourse, and in those notably absent from it. In this context, two critical perspectives arise. First, how faithfully does computational linguistics *actually* adhere to its positivist posture? And second, what are the *implications* of computational linguistics as a singularly positivist discipline? We begin by addressing the former, setting off with a simplified introduction to the positivist worldview and its tenets.

2.1 Number as Virtue

As a scientific meta-theory, positivism asserts that knowledge is the yield of systematic, unbiased and reproducible observation. A prospective theory is evaluated based on how well it can predict and interpret observations. An impartial and irrefutable

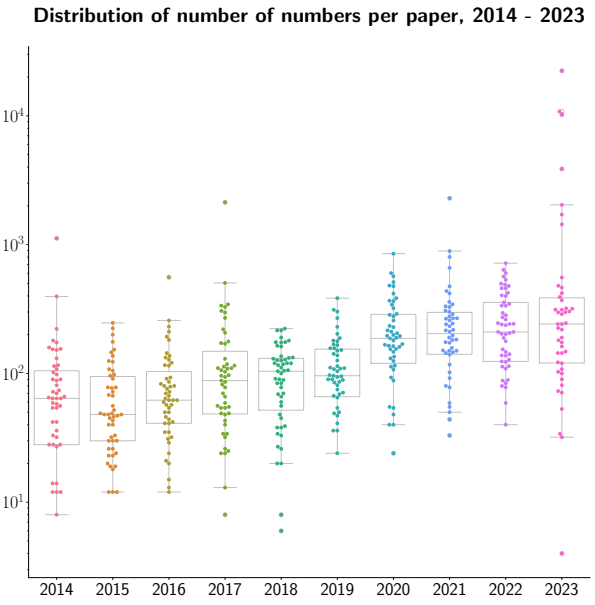

Figure 1: Box- and swarm-plots of the distribution of the number of experimental results per paper, grouped by year. We manually count the number of numbers within tables from the 50 most cited papers per year. We do not include numbers that pertain to descriptive dataset statistics, nor numbers reporting dispersion statistics (*e.g.*, confidence intervals, standard deviations *etc.*). The pattern indicates a marked upwards trend over time. Most (75%) contemporary papers contain 100 to 300 numbers, while some (25%) contain up to 1 000.

evaluation is what ensures theories can be refuted and reliably compared. Ultimately, the essence of scientific progress lies in the iterative process of theory testing, rejection, and refinement. This worldview holds truth as objective and unique, asserted as such by reproducibility, generalization, neutrality, and universality (Ayer, 1959). Tables with numbers attain epistemic significance in bearing witness to this (idealization of) truth.

2.2 Number as Number

Alas, linguistic theories have fallen short of historical expectations. To date, there is no hint of a consensus on what a concretely implementable mechanization of human language should (or even could) look like. In lieu of theories, computational linguistics had to turn to the next best thing: models.[1,2] Models promise less but do more, prioritizing tangible solutions over abstract notions of inquisitive deduction. Apart from this deviation, the positivist methodological narrative is easy to recognize in the field's experimental pipeline. Large datasets are heralded as authoritative collections of empirical observations, systematically condensing linguistic truth. Datasets enact "benchmarks", standardized and fair test suites through which we can "track progress", *i.e.*, decide whether a model advances science, and if so, by *how much*. Congruent with the literature's makeup over the last decade, this suggests that contributions may come in one of two primary forms: models and benchmarks, dual facets of one and the same thing – tables with numbers.

Nonetheless, in having discarded theory, the model-and-benchmark pipeline fails to uphold the scientific promise upon which it was built. A first problem lies in the fact that the models developed and adopted nowadays are almost exclusively generic and theory-neutral (Sutton, 2019). In making no assumptions and yielding no hypotheses over their domain, they are infallible in all aspects except for their performance (Schlangen, 2021). The side effect is that the field's progress translates to technical know-how rather than an advance in the sum total of "pure" knowledge (Krenn et al., 2022; Messeri and Crockett, 2024). Other than modeling insights, nothing gets in and nothing gets out, confining a traditionally interdisciplinary endeavour to a technocratic and opinionless monoculture.

A second, perhaps bigger, problem lies in the reductionist view of language faculty as something that can be broken apart into high-level "tasks", at the intersection of which one can find, and therefore *quantify*, "understanding" (Raji et al., 2021). The verity of this assumption is not immediately obvious; modern models breeze through benchmarks, yet we remain as far as ever from attaining a holistic and comprehensive computational account of language. The picture is sufficiently clear: sidetracked by models and benchmarks, computational linguistics has given way to natural language processing: a domain-specific engineering discipline that is happy to answer more questions than it asks.

2.3 Number as Nothing

Ironically, the remarkable ease of model iteration (as compared to the painstakingly slow process of theory iteration) is an inflationary factor for the epistemic value of numbers. When experimental superiority becomes a prerequisite to publi-

[1]This is one reading. Another reading is that when machine learning "solved" vision, it moved over to NLP, setting aside linguistic expertise to make room for all the luggage it brought with it.

[2]The modern tendency to look for a theory *within* the model (see Baroni (2022); Piantadosi (2023), *inter alia*) is further evidencing the poverty of historical theories.

cation (Rogers, 2020), all publications invariably achieve it, rendering both the message (experimental superiority) and the messenger (publications) meaningless. Immediate, short-sighted gains dominate the research agenda, and difficult questions become eschewed for the sake of incremental tweaks and micro-improvements (Bhattacharya and Packalen, 2020). Short-sighted goals are echoed in short-term memory, leading to plentiful instances of knowledge recycling, paper duplication and citation amnesia (Singh et al., 2023). The over-standardization of form gradually turns into an equilibrium of intent – contributions are pushed towards structural and semantic uniformity, ending up virtually indistinguishable from one another. The frantic pace of "progress" turns scientific enterprise into a competition for experimental superiority, eroding integrity and transparency. The most successful models are too time- and resource-consuming to replicate and cross-validate, leading to statistically insignificant tables filled with under-sampled and noisy numbers of dubious quality and utility (Dodge et al., 2019; Ethayarajh and Jurafsky, 2020; Belz et al., 2021). Scientific communication espouses sales pitch aesthetics, exaggerating merit, obscuring weakness and purposefully avoiding critical self-reflection and honest self-assessment (Smaldino and McElreath, 2016; Lipton and Steinhardt, 2019). After a bountiful decade of benchmarking frenzy, there is now growing consensus that annotation is subjective (Geva et al., 2019; Plank, 2022), datasets are statistically biased, and models are sensitive to heuristics and label noise (McCoy et al., 2019; Geirhos et al., 2020) – the numbers *have been lying all along* (Recht et al., 2019; Liao et al., 2021)![3] Put simply, the more tables with numbers there are, the less a table with numbers means, and the less it can be trusted.

2.4 Number as Vice

Its failure to really adhere to the positivist ethos does not absolve computational linguistics from having adopted it in the first place. The idealization of science as an entity far and above subjective human reference provides the grounds for its disconnect from social context; there's no reflection on its production and consumption, the people involved in it and the people affected by it, or its effect on broader society and the world at large. This detachment is reinforced by a techno-determinist narrative of a "progress" moving of its own accord, which the scientist neither can influence, nor is responsible for (Wyatt, 2008). Tables with numbers are the embodiment of techno-determinism. The quest for experimental superiority (*i.e.*, "progress") is perceived as a self-efficient treadmill that continues on, regardless of who walks it – there's no challenging the pace.

Setting off from a different axiomatization of scientific truth allows for different inference paths. By reflecting on the philosophy of contemporary computational linguistics, we are afforded the opportunity to challenge this particular interpretation of progress – not just for its lack of scientific merit, but more importantly for its active role in perpetuating and amplifying social and environmental harm. We build on this perspective in the following sections.

3 Resource Exhaustion

As the field is witnessing a constant influx of progressively larger models, each vying for supremacy over increasingly more challenging benchmarks, tables are growing in both size and count; see Fig. 1. Meanwhile, the numbers within are getting more resource-intensive by the day (Sharir et al., 2020). As a result, the environmental footprint of contemporary research is expanding at an alarming rate (Strubell et al., 2019; Li et al., 2023).

3.1 No NLP to Be Done on a Dead Planet

The point has resonated with the ecological sensibilities of the community, prompting a number of responses to the issue. By now, these have come to coalesce into a niche of their own, united under the common banner of a so-called "green AI" (Schwartz et al., 2020). So far, most of this green literature has gravitated around two thematic pillars (Verdecchia et al., 2023). The first involves matters of high-level policy: promoting greener models, raising awareness, stamping algorithms and models with eco-labels, *etc.* The second involves matters of low-level practice: truncating or quantizing models, optimizing resource utilization, improving performance-to-emission ratios, *etc.* While both are valuable research avenues, neither really addresses the essence of the problem: the benchmarking practice itself. Indeed, ecologically rooted condemnations of the current *modus*

[3]The fact that benchmarking is being made obsolete by a handful of closed source models far beyond the community's reach is clearly just a coincidence to the timing of this realization.

operandi are rare and far between (with Brevini (2020, 2021, 2022, *inter alia*) and Heilinger et al. (2024) being among the few notable exceptions).

In this case, failing to note the obvious is not (just) a problem of deductive inadequacy; the omission is actually a take in disguise. An ideological child of techno-determinism, on the one hand, and eco-modernism, on the other, it implicitly proclaims that there is no standing in the way of progress – yet *good* progress *can* save the world! The incompatibility of these two positions is glaring. There is little point debating the inherent benevolence of a progress that we cannot contest or control. That said, there is no need to shy away from connecting the dots either. Experimental obsession negatively contributes to a rapidly deteriorating environment, and computational linguistics can never truly be "green" as long as it remains attached to it. The ecologically responsible course of action is not to alleviate the effects – it is to dismantle the cause.

4 Institutional Bias & Privilege

Besides environmental concerns, keeping up with contemporary research trends comes at a (literal) heavy price. As the cost of the "*state of the art*" explodes at a super-exponential rate (Sharir et al., 2020; Epoch AI, 2023; Perrault and Clark, 2024, *inter alia*), the severe budget inequalities in higher education become further pronounced (O'Sullivan, 2016; Goyes and Skilbrei, 2023), and the minimum requirements for scientific relevance becoming prohibitively high for smaller and lesser-funded institutions to acquire and maintain (Ahmed and Wahed, 2020); see also Fig. 2. Consequently, a few dominant institutions get to consolidate their competitive advantage by effectively gatekeeping the means necessary to conduct exactly the kind of research that is perceived as groundbreaking and impactful (Münch, 2014). This is problematic on multiple levels.

4.1 Science of the Few

To begin with, the insurmountable entry barrier perpetuates and exacerbates a cycle of entrenched privilege, where only a few voices retain access to the platforms of expression. This disparity translates the lack of diversity in *what* research is done to a lack of diversity in *who* gets to do it (Ahmed and Wahed, 2020; Perrault and Clark, 2024). For those favored, the cycle is no easier to break. The

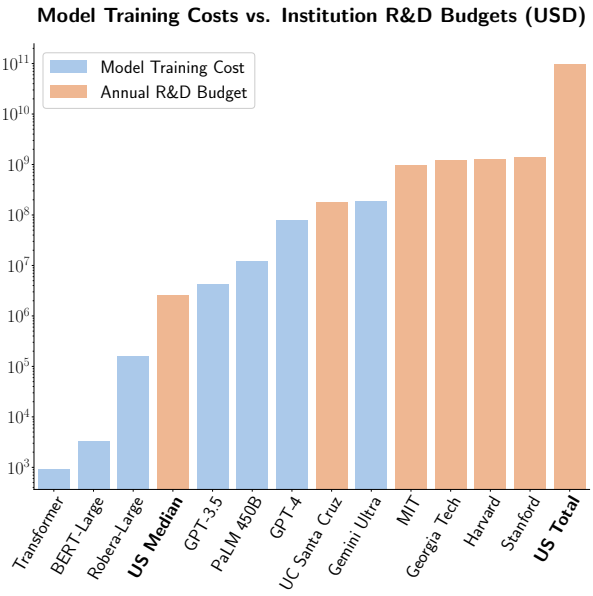

Figure 2: Contemporary model training costs compared to the total annual R&D budgets of select U.S. institutions in 2022. The cost of training a large model is comparable to the budget of a university in the top 15th percentile, which is two orders of magnitude larger than the median budget. Budget data sourced from the 2022 report by the US National Center for Science and Engineering Statistics[a]. Model cost estimates from Epoch AI (2023). The U.S. was the globe's highest spender for the year, in terms of R&D expenditures.

[a] https://ncsesdata.nsf.gov/profiles/

current status quo presents a very alluring prospect: a research recipe that is universally recognized as superior, and that only few have the ingredients necessary to implement. Opting out is not just a matter of critical reflection – it is actually harmful to one's own interests (as measured in publications, citation counts, employment opportunities, *etc.*). Beyond the individual, the same dynamics appear at the institutional scale. Steering a unit away from the competition for experimental superiority and towards niche research means condemning it into academic obscurity and irrelevance; both too easy to mistake for incompetence. This further disincentives scientific plurality, placing the field on a convergent path toward a strict hierarchy of methodologies and ideas, mirrored in a dual hierarchy of institutions and individuals (Rungta et al., 2022).

Effect being all too easy to mistake for cause, a few institutions have by now come to be lauded as hubs of research pioneers, their output singled out and preemptively lauded on the basis of origin alone (Rigney, 2010; Brennen et al., 2019). Privi-

leged individuals are granted undue influence over the field's trajectory, effectively getting to dictate both *what questions to ask* (*e.g.*, which datasets to tackle), and *where to look for the answers* (*e.g.*, which models to adopt). This concentration of technical and scientific authority creates clearly delineated points of vulnerability for the field. Alternative viewpoints and methodologies are at an increased risk of being left unnoticed or becoming squelched, suppressing innovation and inducing inertia. Worse yet, it allows for the biases, norms and opinions of a few dominant actors to be perpetuated unhindered, except now disguised as universal and irrefutable truths characterizing the entire discipline.

4.2 Science for the Few

This last issue is exacerbated exactly by the inherent narrowness of these biases, norms and opinions. Prestigious (read: *wealthy*) institutions are neither evenly distributed across geographic regions, nor equally accessible across social, cultural, ethnic and economic backgrounds. As such, the perspectives and priorities they represent are inevitably skewed towards certain demographics, fostering homogenization at the expense of further marginalizing under-represented groups and identities (Amsler and Bolsmann, 2012; Shamash, 2018; Field et al., 2021; Talat et al., 2022; Hershcovich et al., 2022; Bender and Grissom II, 2024; Perrault and Clark, 2024, *inter alia*). On the premise that cultural diversity is indeed worth nurturing and preserving (Harmon, 2001), the absence of plurality caused by this delegation of scientific and technological authority is bad – for any scientific field. For a field like computational linguistics in particular, it is *catastrophic*. Allowing research agendas to be shaped by a handful of actors endorses hegemonialism: not just technological and scientific, but importantly also cultural and linguistic.

This is particularly evident in the stark geographic disparity between citation-producing networks and centers of linguistic diversity (Rungta et al., 2022). Trending terms like "natural language understanding" carefully conceal the assumptions made on *which* languages are actually worth understanding – or what *understanding* means, for that matter (Bender et al., 2021). The perspective that chasing after benchmarks and competing for the top spots in scoreboards carries some inherent value to the study of language becomes immediately exposed as biased and flawed upon noticing that the majority of benchmarks and scoreboards pertain only to a minuscule fragment of the globe's peoples (Joshi et al., 2020; Ruder, 2022).

Finally, a disproportionate allocation of resources creates the necessary preconditions for scientific tokenism. Technological abundance for the few is indistinguishable from technological sparsity for the many. The surging pressure for inclusivity is temptingly easy to relieve, either by reducing the bar when it comes to work in under-represented languages and cultures, or by "allowing" it to coexist along the mainstream as a secondary, self-referential niche. And while this might indeed expedite its progress or increase its visibility, it carries the risk of negatively impacting its (perceived) quality, further cementing the gap between center and periphery worlds – in terms of language, culture and research alike.

4.3 From Inequality to Alienation

Along the same lines, in monopolizing the resources essential for "frontier" research, "world-class" institutions gain a competitive edge in attracting highly sought-after global talent. Predictably, transnational academic mobility flows along research capacity gradients shaped by global wealth inequalities (Bilecen and van Mol, 2017). The exclusivity of "frontier" research turns academic mobility into a violent dilemma: move, or (academically) perish. Built on this premise, "frontier" research cannot but carry a commodified and socially charged undertone (Stein, 2017).

Two orthogonal aspects of this perspective share a single common effect. First, the same process that accelerates well-funded and globally competitive research decelerates regional institutions and projects by starving them of (yet) another precious resource: talent (Auriol et al., 2013; van der Wende, 2015, *inter alia*). Second, the inherently globalized nature of benchmarking and its constructed significance means that researchers employed abroad are predominantly engaged with work far detached from their own cultural and linguistic heritage. The mirror image of an international researcher pushing the boundaries of "cutting-edge" research is an expatriated researcher not getting their own mother tongue up to speed with that very same research. This reveals benchmarking as a driver for scientific assimilation, which turns linguistic coverage into a matter of institutionalized charity – left to the discretion of exactly those fueling (and benefiting from) its absence.

5 Science & Profit

Albeit alarming, institutional bias is to some extent mitigated by a common (if subjective and vague) promise of scientific integrity, a culture of transparency and openness, a shared strive for intellectual inquiry, and the self-regulatory effect of the (occasionally functional[4]) peer-reviewing system. However, as the race for experimental superiority intensifies, turning increasingly exclusive, each new milestone gains greater appeal. Beyond signaling intellectual achievement or academic accomplishment, this appeal extends to the material plane. There, leading the benchmark race translates to a tangible competitive edge in commercial (and/or state) applications. The allure of such an edge has been persistently attracting profit-driven entities into the computational linguistics ecosystem. Over the span of a decade, these entities have evolved from circumstantial players to dominant figureheads. For such entities, *none* of the safeguards above hold. This reality poses an existential threat for the field; a threat which nonetheless remains largely unaddressed.

5.1 Stand on the Shoulders of (Tech) Giants

The current state of affairs can be traced to a historical affinity between computational linguistics and machine learning (Manning, 2015). Such an affinity is hardly surprising. Language poses challenges at a variety of modalities and difficulty scales, enacting a boundless source of benchmarks for machine learning models. Conversely, models and techniques developed for language-related tasks have frequently demonstrated their versatility as general-purpose machine learning tools, making their way to distant or even unrelated disciplines. Until recently, this reciprocal relationship has been beneficial to both fields. In the last few years, however, and as the pace of progress in machine learning has been consistently exceeding expectations, computational linguistics has lost its primacy, becoming increasingly dependent on imported expertise. This trend is reflected in the silent but perfectly evident shift of the field's main inquiries, which have gradually moved from the computational study of language to an evaluation arena for application-oriented machine learning. And even though this transition might disappoint or alienate some, there is not much inherently wrong about it; after all, it is not uncommon for a research field to retroactively change direction, or even be altogether absorbed or subsumed by another. What *is* problematic in the present context is the nature of the subsumer.

The main pathology of machine learning, having become synonymous with AI, is none other than its public and commercial appeal. The commercialization of science demands tangible advantages against competitors: the product is easier to sell when it's visibly and quantitatively better than alternatives. The success of this commercialization depends largely on "*wow!*" factors: publicity stunts, catchy claims, and a degree of speculative futurism (Funk, 2019). For the global actors invested in the AI race, the concept of performance is thus of prime interest (Bourne, 2024). Current technology dictates one base ingredient as the necessary and sufficient condition for performance: scale (Epoch AI, 2023). And so, we get once more caught up in a vicious cycle. As profit requires performance, performance requires scale, and scale requires budget, a positive feedback loop ensures the growth of a handful of tech giants – at a rate far exceeding that of even the wealthiest research institution. And as performance just so happens to be our currency of choice when quantifying scientific advancement (Birhane et al., 2022), machine learning research becomes *de facto* dominated by exactly these giants (Perrault and Clark, 2024; de Sousa, 2024). This elevates the resource allotment problems discussed earlier to an altogether different scale: what's at stake now is not just equal and fair access to an equal and fair science, but rather the very idea of independent scientific inquiry (Abdalla and Abdalla, 2021; Jurowetzki et al., 2021).

5.2 Research (and Development)

In practice, as long as computational linguistics research remains results-oriented, reliance on technology and infrastructure provisioned by tech giants is a nonchoice – there is, after all, *no one else* to provision them from (Whittaker, 2021; Abdalla et al., 2023; Ferrari, 2023). One might argue that such an arrangement is not without merits. The narrative would usually be that putting corporate technology into the scientific spotlight facilitates the assessment of its risks and potentials, promoting accountability through transparency. Conversely, integrating corporate resources into academia accelerates the actualization of research and increases its impact: rough prototypes turn into concrete tools, ensuring that scientific advancements reach the pub-

[4] See Rogers (2020) and Rogers and Augenstein (2020).

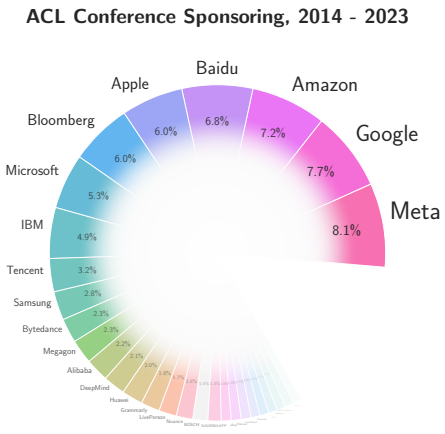

Figure 3: Major sponsors of the main ACL conferences over the last 10 years. To convert tiered participation counts to contributions, we assign a weight of 1 to the year's top tier, and divide the weight of each consecutive sponsorship tier by 2. The treasurer of the ACL did not respond to our request for accurate donation figures.

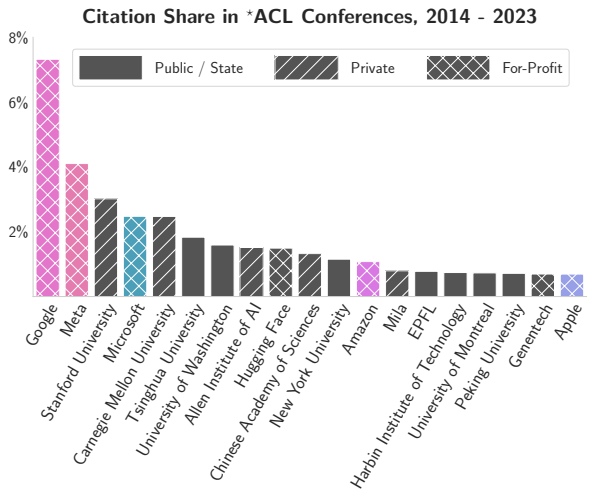

Figure 4: Citation share by organization in *ACL conferences over the last 10 years. Colors are inherited from Fig. 3, when applicable. The 19 organizations listed amount for approximately one third of the total citations during this period. We associate (i) papers to authors, by parsing the ACL bibliography file, (ii) authors to affiliations, by crawling `google scholar` with `scholarly`, and (iii) papers to publication counts, using `Zotero` and the `ZoteroCitationCountsManager` plugin. We collapse affiliations to organizations (*i.e.*, remove job titles and departments) by instructing `mistral-7B` (Jiang et al., 2023). We compose and aggregate over the above to produce a map from organizations to citation counts, disregarding organizations with less than 5 citations as likely parsing errors. The result is imperfect: there are multiple sources of error, and affiliations at retrieval time are likely to differ from those at publication time. Nonetheless, it paints a sufficiently clear picture of which organizations are exerting the most influence in the field, and what the extent of this influence is.

lic domain faster. But such a narrative depends on, and in fact presupposes, an alignment between scientific and commercial agendas. The implication is that the pursuit of knowledge becomes conditional on its compatibility with the interests and capabilities of big tech, *i.e.*, the very same actors academia was supposed to scrutinize in the first place.

The conflict of interest is immediately apparent. The overwhelming power asymmetry between big tech and academia (be it big or otherwise) erodes any potential merits that could ever be argued for. Under the present conditions, the scientific spotlight can no longer be critical or investigative. Conferencing devolves to a campaigning stage, a ticketed tech show, and a marketplace where for the colossi to display their latest wares and recruit new talent; see Fig. 3, and juxtapose with Fig. 4. Corporate resources do not "spill over", nor do they "trickle down" – they are rationed; a means of scientific coercion (Noble, 1979; Moore et al., 2011; Phan et al., 2022). Corporate interests do not actualize knowledge – they predate, appropriate and monetize it (Rikap and Lundvall, 2022). Ideas that survive the ecosystem's selection process do not turn into socially relevant tools – they turn into economically viable products (Dale, 2019; Klinger et al., 2020; Luitse and Denkena, 2021). Scientific involvement itself degrades into a "networking filter": an inconvenient but unavoidable stepping stone towards a high-stakes career in tech (Ahmed et al., 2023; Gofman and Jin, 2024). The researcher becomes a glorified spokesperson for big tech, a consumer of their infrastructure, a public advocate of their science, a safety net between them and the public – an eager and dispensable part of their production pipeline.

The extent and degree of the infiltration have become impossible to ignore. We are on the verge of a corporate takeover, legitimized by an acquired taste for big datasets, big models and big numbers. Put simply, we have been voluntarily handing the field over to an industry we are realistically incapable of challenging, let alone regulating.

5.3 (The Irrelevance of) Corporate Ethics

As of late, the community's growing awareness (Michael et al., 2023) of these developments and their public ramifications has spurred numerous works on so-called "AI ethics". The conversation is heavily skewed by well-documented lobbying efforts and a broader ethics-washing campaign aimed at soothing public concern and deter-

ring regulatory oversight. The "debate" often revolves around virtue signaling gestures, assertions of corporate responsibility (or accusations of its absence), suggestions for self-regulatory accountability guidelines, techno-positive musings of an all-inclusive tomorrow, "critical" perspectives from within, vague calls for a misconstrued "democratization", and the like. In their majority, these works range from malicious manipulation at worst, to harmful diversions at best (Ochigame, 2019; Benkler, 2019; Slee, 2020; Hagendorff, 2020; Whittaker, 2021; Phan et al., 2022; Seele and Schultz, 2022; Himmelreich, 2023, *inter alia*).

This premeditated and narrow notion of ethics subtly chooses to ignore the possibility of us reappropriating the scientific discourse. Besides negotiating matters of representation and inclusion, bias aversion, model explainability, linguistic diversity, open-sourcing, carbon impact, *etc.* as they arise within the *current* environment, we have a far more fundamental series of questions to be confronted with. Are we assuming that big tech, running rampant on the field's collective advancements, will (or even can) ever align their agenda with the public's interests? Do we trust them with upholding the values of scientific integrity and technological accountability? Are we at peace with the prospect of a privatized and application-centric future for computational linguistics, removed from the world, its people and their needs? If the answer to the above is no, how can we justify our implicit yet unwavering support and commitment to big tech's cause throughout the last decade? Why are we so susceptible to their influence, so eager to adopt their values and principles, so tolerant of their technologically exclusionary practices? Ultimately, what benefits do *we* get to derive from contributing to *their* endeavors – and at what cost?

6 Ways Ahead

The paradigm shift advocated for might seem radical or untenable. In reality, it is neither. The epistemic rewiring it calls for can be set in motion with as little as individual adjustments in research consumption and production attitudes.

As *readers*, we need to stop allowing ourselves to be dazzled by big numbers. We must ask what their utility and cost are, who benefits from them, and who bears their expense. We should not only grow resilient to hollow benchmarking hypes, but also openly refute and disarm them.

As *authors*, *colleagues* and *advisors*, we have to be conscious of our (and each other's) research goals and practices. We ought to look beyond numbers and benchmarks and focus on what questions our research really answers. We must challenge the notion of science as a competition or enterprise, and scorn endeavors that depend solely on experimental superiority to be deemed successful. We must be mindful and explicit of the resources we use and their accessibility, but also of the artifacts we produce and their inclusivity. Above all, it is our responsibility to be vocal and assertive about the issues in our field; despite –or rather *in spite of*– normative resistance and calls for conformity and "moderation".

As *reviewers*, we should each recognize our respective academic privileges, and be cautious in our technical demands; not everyone has access to the same number of GPUs. Conversely, we should not be intimidated by big tables and bold face fonts; we need to be critical of the research we are exposed to, and call out opaque methodologies, exclusionary practices and useless flourishes. Finally, our exclusive access to the reviewing process means it is our own duty to monitor it; each one of us has a role in identifying and confronting poor practices.

7 Conclusion

We discussed tables with numbers, and related them to several issues that affect contemporary computational linguistics research. We argued that the focus on experimental superiority has shifted research priorities towards technical optimization, at the expense of theoretical depth and societal context. This has led to an inflationary effect on the epistemic value of experimental results, rendering them (and, by extension, the field itself), increasingly meaningless. We explained how the pressure for experimental superiority, while advancing technology, has fostered environmental degradation, institutional biases, and the commodification of research. To address these issues, we urge the field to critically reassess its methodologies, and prioritize a more holistic and socially responsible approach to scientific inquiry, balancing technical achievements with ethical and environmental considerations. Such a shift is essential for ensuring that advancements in computational linguistics positively contribute to scientific knowledge, societal well-being, cultural diversity, and environmental sustainability.

Limitations

We tried to substantiate our claims with (references to) empirical evidence and contemporary critical perspectives. Nonetheless, this paper is first and foremost an opinion piece; the ideas presented are the product of subjective and ideologically signed mental processes. For a reader that ascribes to the epistemic foundations of positivism, this is an argumentative weakness. For us, it is a strength. We acknowledge our biases and limitations, and welcome critiques from all angles; a broader discussion on the field's epistemic culture is exactly what our work hopes to instigate.

Our analysis is by no means exhaustive, especially considering the complexity and volatility of the subject matter. The most critical omission, due to the temporal gap between writing this piece (August 2024) and getting it published (March 2025), is a reflection on how recent political developments have further validated the transient and opportunistic nature of big tech's so-called ethics. Following the change in power after the USA 2024 elections, tech companies have been increasing their stakes in transnational military and surveillance applications, while simultaneously backpedaling on their own commitments on ecological sustainability and social diversity, equity and inclusion. We defer a discussion on the military-industrial complex emerging from key players in the language technology industry for another occasion.

Finally, there are several experiments we would have hoped to carry out to quantify some of our claims, but we failed to bring to fruition. We explicitly mention them here for the sake of clarity and transparency, and to bring them to the attention of other interested parties:

- A paper-wise computational cost estimation would allow a quantification of the financial entry barrier to modern research. Overlaid with citation counts, this would allow answering whether the most impactful papers are really just the most expensive ones.
- A longitudinal topic modeling analysis could provide evidence for the narrowing of research topics and methodologies over the last decade. Combined with an evolutionary analysis of writing norms (e.g., paper structure), this would allow us to correlate homogenization of tone with the loss of content diversity.

Acknowledgments

The second author is partially funded by the European Union (ERC ADG, PhylProGramm, 101096554). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

The second author is also partially funded by the Special Account for Research Funding of the Technical University of Crete (grant number: 11218).

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*, pages 287–297.

Mohamed Abdalla, Jan Philip Wahle, Terry Ruas, Aurelie Neveol, Fanny Ducel, Saif Mohammad, and Karën Fort. 2023. 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)*, pages 13141–13160.

Nur Ahmed and Muntasir Wahed. 2020. The de-democratization of AI: Deep learning and the compute divide in artificial intelligence research. *arXiv preprint arXiv:2010.15581*.

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

Sarah Amsler and Chris Bolsmann. 2012. University ranking as social exclusion. *British journal of sociology of education*, 33(2):283–301.

Laudeline Auriol, Max Misu, and Rebecca Ann Freeman. 2013. Careers of doctorate holders: Analysis of labour market and mobility indicators. *OECD Science, Technology and Industry Working Papers*.

Alfred Jules Ayer. 1959. *Logical positivism*, volume 2. Simon and Schuster.

Marco Baroni. 2022. On the proper role of linguistically oriented deep net analysis in linguistic theorising. In *Algebraic structures in natural language*, pages 1–16. CRC Press.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, Online. Association for Computational Linguistics.

Emily Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Emily Bender and Alvin Grissom II. 2024. Power shift: Toward inclusive natural language processing. *Inclusion in Linguistics*, page 199.

Yochai Benkler. 2019. Don't let industry write the rules for AI. *Nature*, 569(7754):161–162.

Jay Bhattacharya and Mikko Packalen. 2020. Stagnation and scientific incentives. Technical report, National Bureau of Economic Research.

Başak Bilecen and Christof van Mol. 2017. Introduction: International academic mobility and inequalities.

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*, pages 173–184.

Clea Bourne. 2024. AI hype, promotional culture, and affective capitalism. *AI and Ethics*, pages 1–13.

Scott Brennen, Anne Schulz, Philip Howard, and Rasmus Kleis Nielsen. 2019. Industry, experts, or industry experts? Academic sourcing in news coverage of AI. *Reuters Institute for the Study of Journalism*.

Benedetta Brevini. 2020. Black boxes, not green: Mythologizing artificial intelligence and omitting the environment. *Big Data & Society*, 7(2):2053951720935141.

Benedetta Brevini. 2021. *Is AI good for the planet?* John Wiley & Sons.

Benedetta Brevini. 2022. Dispelling the 'green' AI myth: The true environmental cost of producing and supplying digital technologies. Accessed: 26-06-2024.

Kenneth Church and Mark Liberman. 2021. The future of computational linguistics: On beyond alchemy. *Frontiers in Artificial Intelligence*, 4:625341.

Robert Dale. 2019. NLP commercialisation in the last 25 years. *Natural Language Engineering*, 25(3):419–426.

Miguel Angelo de Abreu de Sousa. 2024. The shift of artificial intelligence research from academia to industry: Implications and possible future directions. *AI & SOCIETY*, pages 1–10.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Epoch AI. 2023. Key trends and figures in machine learning. Accessed: 2024-05-29.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853.

Fabian Ferrari. 2023. Neural production networks: AI's infrastructural geographies. *Environment and Planning F*, 2(4):459–476.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925.

Jeffrey Funk. 2019. What's behind technological hype? *Issues in Science and Technology*, 36(1):36–42.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Michael Gofman and Zhao Jin. 2024. Artificial intelligence, education, and entrepreneurship. *The Journal of Finance*, 79(1):631–667.

David Rodriguez Goyes and May-Len Skilbrei. 2023. Rich scholar, poor scholar: Inequalities in research capacity, "knowledge" abysses, and the value of unconventional approaches to research. *Crime, Law and Social Change*, pages 1–20.

Sireesh Gururaja, Amanda Bertsch, Clara Na, David Widder, and Emma Strubell. 2023. To build our future, we must know our past: Contextualizing paradigm shifts in natural language processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13310–13325, Singapore. Association for Computational Linguistics.

Thilo Hagendorff. 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds and machines*, 30(1):99–120.

David Harmon. 2001. On the meaning and moral imperative of diversity. In Luisa Maffi, editor, *On Biocultural Diversity: Linking Language, Knowledge, and the Environment*, pages 53–70. Smithsonian Institution Press.

Jan-Christoph Heilinger, Hendrik Kempt, and Saskia Nagel. 2024. Beware of sustainable AI! Uses and abuses of a worthy goal. *AI and Ethics*, 4(2):201–212.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, et al. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013.

Johannes Himmelreich. 2023. Against "democratizing AI". *AI & SOCIETY*, 38(4):1333–1346.

Albert Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Roman Jurowetzki, Daniel Hain, Juan Mateos-Garcia, and Konstantinos Stathoulopoulos. 2021. The privatization of AI research(-ers): Causes and potential consequences. *arXiv preprint arXiv:2102.01648*.

Joel Klinger, Juan Mateos-Garcia, and Konstantinos Stathoulopoulos. 2020. A narrowing of AI research? *arXiv preprint arXiv:2009.10385*.

Mario Krenn, Robert Pollice, Si Yue Guo, Matteo Aldeghi, Alba Cervera-Lierta, Pascal Friederich, Gabriel dos Passos Gomes, Florian Häse, Adrian Jinich, AkshatKumar Nigam, et al. 2022. On scientific understanding with artificial intelligence. *Nature Reviews Physics*, 4(12):761–769.

Pengfei Li, Jianyi Yang, Mohammad A Islam, and Shaolei Ren. 2023. Making AI less" thirsty": Uncovering and addressing the secret water footprint of AI models. *arXiv preprint arXiv:2304.03271*.

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. 2021. Are we learning yet? A meta review of evaluation failures across machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Zachary Lipton and Jacob Steinhardt. 2019. Troubling trends in machine learning scholarship: Some ML papers suffer from flaws that could mislead the public and stymie future research. *Queue*, 17(1):45–77.

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):20539517211047734.

Christopher Manning. 2015. Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Lisa Messeri and M.J. Crockett. 2024. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58.

Julian Michael, Ari Holtzman, Alicia Parrish, Aaron Mueller, Alex Wang, Angelica Chen, Divyam Madaan, Nikita Nangia, Richard Yuanzhe Pang, Jason Phang, et al. 2023. What do NLP researchers believe? Results of the NLP community metasurvey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16334–16368.

Kelly Moore, Daniel Lee Kleinman, David Hess, and Scott Frickel. 2011. Science and neoliberal globalization: A political sociological approach. *Theory and Society*, 40:505–532.

Richard Münch. 2014. *Academic capitalism: Universities in the global struggle for excellence*. Routledge.

David Noble. 1979. *America by design: Science, technology, and the rise of corporate capitalism*. 588. Oxford University Press, USA.

Rodrigo Ochigame. 2019. The invention of 'ethical AI': How big tech manipulates academia to avoid regulation. *Economies of virtue*, 49.

Michael O'Sullivan. 2016. *Academic barbarism, universities, and inequality*. Springer.

Ray Perrault and Jack Clark. 2024. Artificial intelligence index report 2024. *Human-Centered Artificial Intelligence*.

Thao Phan, Jake Goldenfein, Monique Mann, and Declan Kuch. 2022. Economies of virtue: The circulation of 'ethics' in big tech. *Science as culture*, 31(1):121–135.

Steven Piantadosi. 2023. Modern language models refute Chomsky's approach to language. *Lingbuzz Preprint, lingbuzz*, 7180.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682.

Inioluwa Deborah Raji, Emily Denton, Emily M Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR.

Daniel Rigney. 2010. *The Matthew effect: How advantage begets further advantage*. Columbia University Press.

Cecilia Rikap and Bengt-Åke Lundvall. 2022. Big tech, knowledge predation and the implications for development. *Innovation and Development*, 12(3):389–416.

Anna Rogers. 2020. Peer review in NLP: reject-if-not-SOTA.

Anna Rogers and Isabelle Augenstein. 2020. What can we do to improve peer review in NLP? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1256–1262.

Sebastian Ruder. 2022. The State of Multilingual AI. `http://ruder.io/state-of-multilingual-ai/`.

Mukund Rungta, Janvijay Singh, Saif M. Mohammad, and Diyi Yang. 2022. Geographic citation gaps in NLP research. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1371–1383, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Schlangen. 2021. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674, Online. Association for Computational Linguistics.

Roy Schwartz, Jesse Dodge, Noah Smith, and Oren Etzioni. 2020. Green AI. *Communications of the ACM*, 63(12):54–63.

Peter Seele and Mario Schultz. 2022. From greenwashing to machinewashing: A model and future directions derived from reasoning by analogy. *Journal of Business Ethics*, 178(4):1063–1089.

Rebecca Shamash. 2018. (Re)production of the contemporary elite through higher education: A review of critical scholarship. *Berkeley Review of Education*, 8(1).

Or Sharir, Barak Peleg, and Yoav Shoham. 2020. The cost of training NLP models: A concise overview. *arXiv preprint arXiv:2004.08900*.

Janvijay Singh, Mukund Rungta, Diyi Yang, and Saif Mohammad. 2023. Forgotten knowledge: Examining the citational amnesia in NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6192–6208, Toronto, Canada. Association for Computational Linguistics.

Tom Slee. 2020. The incompatible incentives of private-sector AI. *The Oxford Handbook of Ethics of AI*, pages 106–123.

Paul Smaldino and Richard McElreath. 2016. The natural selection of bad science. *Royal Society open science*, 3(9):160384.

Sharon Stein. 2017. Internationalization for an uncertain future: Tensions, paradoxes, and possibilities. *The Review of Higher Education*, 41(1):3–32.

Emma Strubell, Ananya Ganesh, and Andrew Mccallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650.

Richard Sutton. 2019. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics.

Marijk van der Wende. 2015. International academic mobility: Towards a concentration of the minds in Europe. *European review*, 23(S1):S70–S88.

Roberto Verdecchia, June Sallou, and Luís Cruz. 2023. A systematic review of green AI. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(4):e1507.

Meredith Whittaker. 2021. The steep cost of capture. *Interactions*, 28(6):50–55.

Sally Wyatt. 2008. Technological determinism is dead; long live technological determinism. *The handbook of science and technology studies*, 3:165–180.