Resource Democratization: Is Compute the Binding Constraint on AI Research?

Rebecca Gelles, Veronica Kinoshita, Micah Musser, James Dunham

Center for Security and Emerging Technology, Georgetown University
{rebecca.gelles, ronnie.kinoshita, mrm311, james.dunham}@georgetown.edu

Abstract

Access to compute is widely viewed as a primary barrier to AI research progress. Compute resource disparities between academic and industry researchers is therefore a source of concern. Yet the experiences of researchers who might encounter resource constraints in their work have received no direct study. We addressed this gap by conducting a large survey of U.S. AI researchers that posed questions about project inputs, outcomes, and challenges. Contrary to popular narratives, responses from more than 500 participants revealed more concern about talent and data limitations than compute access. There were few differences between academic and industry researchers in this regard. The exception were researchers who already use large amounts of compute, and expressed a need for more. These findings suggest that interventions to subsidize compute without addressing the limitations on talent and data availability reported by our respondents might cause or exacerbate commonly cited resource inequalities, with unknown impact on the future of equitable research.

Introduction

Inequitable access to compute could throw trustworthy and innovative artificial intelligence development off balance. Recent proposals for a National AI Research Resource (NAIRR) in the United States stake $2.25 billion on the belief that compute access is of paramount importance to AI progress. However, an essential voice has been left out of the conversation of the relative importance of data, talent, and compute in AI progress: AI researchers themselves. Compute-heavy development advocates frequently emphasize inequities around large "cutting-edge" deep learning models' compute demands, but do these barriers and compute restraints reflect the concerns of a vocal minority or the experience of the broader community of AI researchers?

Our research contributes evidence to answer these questions by asking AI researchers in the U.S. about their compute usage, concern about future compute access, and the extent to which compute—as compared to factors like data or talent—affects their projects. We find that compute is not the primary constraint faced by most AI researchers, and access to data or talent more directly limits research plans and researcher behavior. We also find little evidence that industry researchers even use significantly more compute than researchers in academia or that academic researchers are more concerned about their access to compute. Our participants' observations regarding the concept of a NAIRR are generally supportive, especially with regard to grant funding, but they cite concerns about ineffective implementation.

Background

Inequity in access to computing resources for AI research is a growing concern. Previous researchers have warned that an increased concentration of computational resources in the hands of a few organizations, mostly industry firms and elite universities, may decrease the overall diversity of the field, in terms of researcher participation, topics studied (Ahmed and Wahed 2020), and worldwide involvement (Depp 2022). Recent research has claimed that publications from elite universities in top AI venues have crowded out researchers from less prestigious universities since 2012, in part due to disparities in compute access (Ahmed and Wahed 2020). Indications of less diversity and increased technical concentration have been observed in the academic literature (Klinger, Mateos-Garcia, and Stathoulopoulos 2020), and AI researchers' career paths: the propensity of top deep learning talents in academia to transition to industry (Jurowetzki et al. 2021). Fields of study like large generative models have a particularly high computational barrier to entry (Ganguli et al. 2022), although in a number of cases researchers have later been able to reproduce work by industry labs at lower cost or using fewer parameters (Izsak, Berchansky, and Levy 2021; Ahdritz et al. 2022; Du et al. 2022; Ding et al. 2022; Dey et al. 2023). Even top generative-AI labs have suggested that expanding computational resources may not be the path to future progress (Knight 2023).

Previous literature review (Paleyes, Urma, and Lawrence 2022) and interview studies (Merhi 2023; Westenberger, Schuler, and Schlegel 2022; Baier, Jöhren, and Seebacher 2019; Shankar et al. 2022) have focused on AI project's success or failure in practice. These studies have unearthed many barriers, both technical and non-technical (e.g., organizational buy-in, ethics, bureaucratic constraints). In the technical realm, lack of quality data and talented practitioners are consistent themes. Compute resources were mentioned much less frequently, although the high cost of AI

was often discussed, and these factors may be connected. These studies focused on the use of AI tools by industry practitioners, rather than on novel research in AI.

More general surveys and interview studies focused on the future of the field of AI have also gathered relevant data from AI researchers and practitioners. While these studies primarily investigated how AI will evolve (Zhang et al. 2022), some also discussed the resources needed to pursue future work and the factors likely to drive progress. Müller and Bostrom (2016) asked respondents what might contribute most to developing high-level machine intelligence (HLMI), such as algorithmic approaches, computing hardware, and data improvements; they found some algorithmic approaches ranked the highest, followed by computation and then data. Gruetzemacher, Paradice, and Lee (2019) surveyed experts about the computational resources needed to pursue future AI progress, finding that the majority predicted a 50% or greater increase in progress with unlimited compute; no comparable question was asked about data or talent. A recent survey, with full results yet unpublished, on AI progress and HLMI included questions about how hypothetical changes in the historical availability of researchers, data, compute, funding and algorithms would have affected overall AI capabilities today; its final results will be informative (Stein-Perlman, Weinstein-Raun, and Grace 2022). Industry firm studies have focused more on AI practitioners than on researchers; their insights remain relevant to machine learning (ML) infrastructure costs and budget growth (Algorithmia; Dotscience); challenges regarding training data (Dimensional Research); and internal compute, data, and talent decisions (McKinsey). One exception to this focus is work describing trends in AI research, which emphasizes the cost of compute and how it may constrain academic research (Benaich and Hogarth 2022).

Another technique to evaluate AI resource constraints has been to measure the cost of building state-of-the-art models. Sharir, Peleg, and Shoham (2020) consider NLP training costs and how they are affected by a variety of model design factors, observing that these costs have increased even as hardware costs have fallen. Similar work finds that the amount of compute used as of 2018 in the largest AI models was increasing exponentially with a 3.4-month doubling time (OpenAI); that compute power has been a major driver in improvement on a variety of AI benchmarks (Thompson, Ge, and Manso 2022); and that model size and compute budget increases alone can dramatically affect model performance (Kaplan et al. 2020). On the other hand, Hernandez and Brown (2020) find strong evidence that AI algorithms are growing significantly more compute-efficient, and Bartoldson, Kailkhura, and Blalock (2022) discuss numerous techniques for building more compute-efficient models. The need for these advances may exist regardless of resource constraints, based on the work of Thompson et al. (2020), which indicates that if compute use continues expanding at the current pace the environmental and economic costs will become unrealistic. More optimistically, Patterson et al. (2022) find that alternative model architecture choices can reduce costs and energy consumption, perhaps mitigating these future challenges.

At this time, no studies have directly asked AI researchers themselves about their resource constraints, or compared the extent of those constraints among different subgroups to better understand potential inequities. We address this gap by applying survey research methods, asking AI researchers about how resource constraints affect their work, and analyzing the results.

Methodology

We asked respondents about their AI projects, compute usage, perspectives on research resources, and opinions regarding the importance of various factors for AI research progress. The survey included 30-35 close-ended questions and one open-ended question, depending on each respondent's employment experiences and AI projects. Early versions of the survey instrument were refined through a series of cognitive interviews with AI researchers in academia and industry. The full instrument is available in an online appendix [1]. The study was approved by the Georgetown University IRB, and granted exemption signifying the research activities were of minimal risk to participants. Participants were prompted to indicate consent, which described the purpose of the research, any risks, and how data would be collected and reported anonymously.

We created a sampling frame by enumerating individuals who authored a paper at a top AI conference or journal, or worked in industry in an AI-related role. We identified authors of papers in 20 leading AI journals or conferences between 2016 and 2021 using Web of Science (see appendix for the list). This resulted in 27,172 authors with email contact information who were affiliated with a U.S. institution at the time of paper publication. Second, we identified industry AI researchers using LinkedIn data from Revelio Labs. We looked for LinkedIn users who 1) reported working as a ML or AI engineer (or similar), or 2) identified their employer as one of 46 AI startups[2] and their role as technical (see appendix for included job titles). We randomly selected roughly 5,000 profiles that met this criteria and used RocketReach, an email sourcing vendor, alongside manual searching to identify emails for 3,894 industry AI researchers.

In total, we received 410 complete responses and 123 partially complete responses, which were also included. We screened respondents at the start of the survey to ensure that they "build, develop, study, or maintain" AI systems "at least some of the time." Three pilot versions of the survey were sent to random samples of 500 identified AI researchers in late spring 2022. We used these pilot surveys to estimate a likely response rate and evaluate response options; we made no substantive changes to the final instrument. Primary survey distribution occurred in June 2022, with a follow-up in July 2022 to 50 researchers who previously had invalid emails. Responses from the pilot and follow-up distributions are included in the analysis. The median survey response time was eight minutes.

[1] Available at https://github.com/georgetown-cset/Compute_Survey_2022

[2] We drew the companies from a 2021 CB Insights report, "The United States of Artificial Intelligence Startups."

Results

Of our 410 complete responses, 274 (67%) report working in academia, 120 (29%) in industry, and 14 (3%) in government. Among respondents who reported working in industry, 84 reported working for a company with over 500 employees, nine reported a company size of 101-500, nine reported a company size of 50-100, and 17 reported a company size of fewer than 50 employees.

To help understand our sample of academic respondents, we looked at the email domains for all AI researchers who were invited, started, or completed the survey. That set included 423 ".edu" email domains: 147 (35%) from a top 50 university, 115 (27%) from a university ranked 51–200, 134 (32%) from a university ranked below 200, and 27 (6%) from unranked universities, according to QS World University Rankings (QS Top Universities). This suggests our sample includes researchers working in a variety of academic institutions. Respondents were also asked to indicate which fields they worked in. Top-level nonexclusive options were computer vision (n = 151), natural language processing (n = 143), reinforcement learning (n = 80), robotics (n = 72), and other (n = 160). A larger share of academics report working in robotics and reinforcement learning, while among industry respondents, a large share report working in natural language processing. Full breakdowns of the number of respondents by field, subfield, and sector can be found in the GitHub repository (https://github.com/georgetown-cset/Compute_Survey_2022). Appendix contains comparisons between subfields across each of the five top-level categories.

Compute Is Not the Primary Constraint for Many AI Researchers

To understand how AI researchers see compute as a driving or constraining resource, we asked a variety of questions, prompting respondents to indicate the relative importance of compute, data, and talent for their projects, their resource priorities given a larger budget, how often resources altered project plans, and the importance of compute in driving AI progress to date and in the future.

Finding 1.1. Researchers Report Talent As the Primary Factor Contributing to the Success of Their Most Significant Projects and Most Researchers Would Prioritize Talent if Given More Funding. To evaluate resource constraints, we asked respondents about two projects they worked on in the previous five years: the one they believed made the "most significant contribution" to research progress in their field, and their "most compute-intensive" project. For 67% of respondents these were the same project; despite this, we found researchers rate talent as more important to their most significant project's success. 90% of respondents rated "specialized knowledge, talent, or skills" as very or extremely important for said project's success, compared to 52% rating "large amounts of compute" as similarly important (see Figure 1). A similar proportion (51%) rated "unique data" as very or extremely important. This question asked respondents to rate each factor independently, but other questions asked respondents to compare compute to other factors, and talent again surfaced as an important resource.

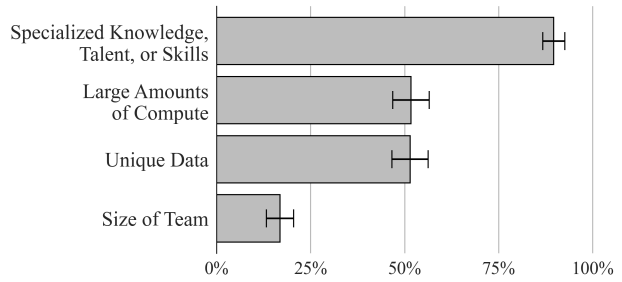

Figure 1: Percent of respondents viewing factors as important for project success

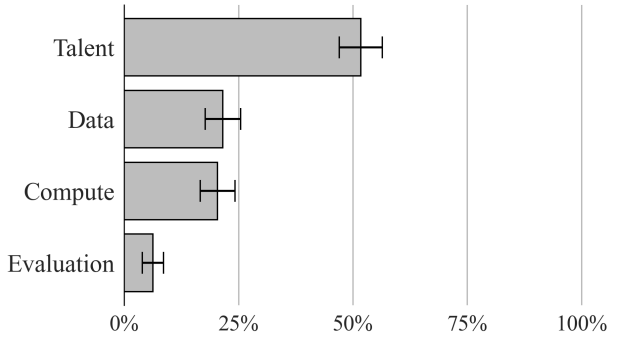

Figure 2: Percent of respondents selecting factors as their top budget priority

To assess how researchers prioritize resources, we asked them to imagine the budget for their current or most recent AI project doubled: what would their first spending priority be? Roughly half (52%) said that they would first spend the additional money on either "hiring researchers" or "hiring more programmers or engineers," which we display together in Figure 2 under "Talent." About a fifth of researchers would make "purchasing more or higher-quality compute" their first priority and a similar share would first use the funds to collect or clean data.

Finding 1.2. When Researchers Are Forced To Change Their Research Plans, It Is More Often Due to Talent or Data Limitations Than Compute Limitations. One indication that a resource is a constraint on progress is that insufficient access to that factor affects research plans. To explore this possibility, we asked how often over the past two years respondents rejected, revised, and abandoned an ongoing project due to insufficient compute, data, and researcher availability. The mean responses on a five-point Likert scale are shown in Figure 3.

Researchers report rejecting and abandoning projects due to a lack of data or researcher availability more often than from insufficient compute resources. Researchers also report revising ongoing projects more often due to data constraints (but not a lack of researcher availability), as compared to a lack of compute resources. All pairwise comparisons were

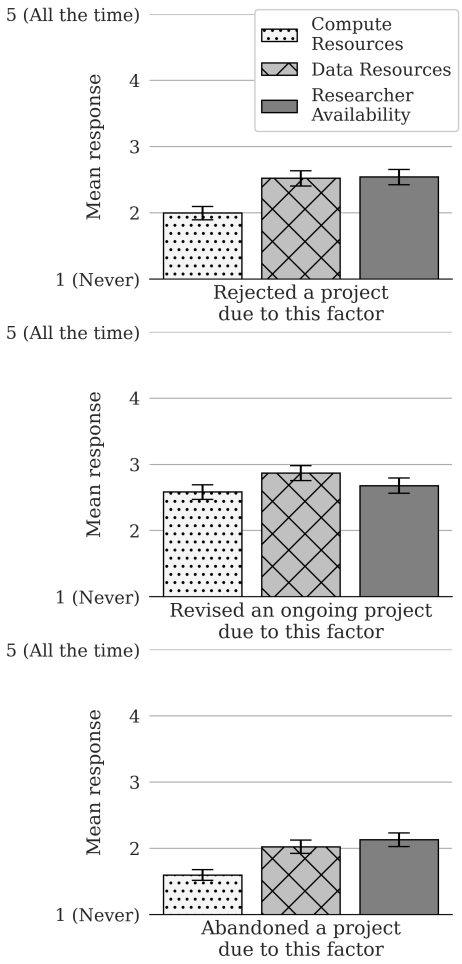

Figure 3: Rates at which respondents change research plans due to various factors

made using Mann-Whitney U tests with the Bonferroni correction. Differences between compute and data were significant for rejecting, revising, and abandoning projects (all $p < .001$). Differences between compute and researcher availability were significant for rejecting and abandoning projects (both $p < .001$) but not for revising projects (p = .773). Differences between data and researcher availability were significant for revising projects (p = .038) but not for rejecting (p = 1.0) or abandoning projects (p = .497).

While constraints from data and talent are more often reported as reasons for rejecting or abandoning a project, 76% of respondents report revising projects due to insufficient compute at least "sometimes" over the last two years. However, 43% of respondents report never rejecting a project due to insufficient compute, indicating that some subset of AI researchers are able to pursue the research they want at their current level of compute resourcing.

Finding 1.3 Most Respondents Think Computing's Role in Driving AI Progress Will Stay the Same or Decrease in the Next Decade, Compared to the Past Decade. To evaluate compute's relative value in past and future progress, we asked respondents for their level of agreement with the claim that progress in AI over the past decade was the result of five different factors: data, compute, algorithms, number of researchers, and level of support for AI projects. Respondents agreed that each factor contributed to past AI progress, with 59%, the highest level of agreement, attributing past success to compute.

Fewer respondents (40%) strongly agreed that compute would be a driver of AI progress over the next decade. The largest gap between retrospective and prospective evaluations among the factors was with better algorithms, which 31% strongly agreed had driven past AI progress, but 53% strongly agreed would drive future progress. Table 1 shows the change in strong agreement for each factor's influence on past and future AI progress.

Factor	% Strongly agree AI progress was or will result from this factor		Difference
	Past decade	Next decade	
Better algorithms	31%	53%	+22%*
Greater support for AI projects	33%	42%	+9%
More researchers in the field	32%	35%	+3%
More or better data	47%	44%	-3%
More compute	59%	40%	-19%*

Table 1: Perceived importance of factors for AI progress[3]

Reported Compute Use Is Similar for Industry and Academia

To examine differences in compute use and needs between academic and industry researchers, we break down responses to various questions according to the respondent's reported employment in academia or industry; government researchers were omitted due to sample size.

Finding 2.1. Academics Report Paying Less for Compute but Do Not Report Significantly Less Compute Use. To understand the variability in industry and academia researchers' maximum compute access and need level, we asked several questions about the most compute-intensive AI project that they had worked on in the preceding five

[3]Asterisks indicate statistically significant differences ($p < .001$) in response distribution for a factor over the last decade compared with response distribution for that factor over the next decade using a Mann-Whitney U test with Bonferroni correction.

years. When asked how expensive the compute required by this project was, academics reported spending significantly less ($p < .001$ by a Mann-Whitney U test) than industry researchers. This finding is consistent with the narrative that the compute capabilities of industry researchers are rapidly outpacing those of their academic counterparts. When asked about compute use for this project in terms of GPU-hours, however, we observe no meaningful difference ($p = .200$). See appendix for two ordinal logistic regression models providing further analysis.

While we find no reported difference in compute use, as measured by GPU-hours, for their most significant project, this does not capture all possible differences in compute access between industry and academic researchers. For instance, one possibility not covered here is that academics may be using cheaper—and lower-performing—GPUs than those used in industry. We nonetheless regard GPU-hours as the better measure for compute use for several reasons. First, 332 respondents provided information about GPU-hours, compared to only 261 respondents for cost.[4] While some might argue dollars are the more salient metric, the fact that more researchers are able to report GPU-hours than dollars suggests this may not be the case.

Second, some researchers who use on-premise compute - which has already been paid for - may report "$0" and significantly ($p < .001$ by a chi-squared test of independence) more on-premise users are academics, at 82% (46% exclusively) as compared to only 52% (22% exclusively) of industry researchers. Third, cloud computing companies often provide access to compute resources at discounted rates for academics. Combined, these factors make monetary cost a less reliable measure of compute use across sectors.

Finding 2.2. Academics Report Compute Needs Have Outpaced Availability but Are Not Significantly More Concerned About Future Access Impacting Their Contributions to AI. We then directly asked respondents how much compute they need and how much they have access to, relative to two years ago; results are shown in Figure 4. We observe a significantly ($p = .004$, by a chi-squared test of independence) greater proportion of respondents in academia, as compared to industry, report that their change in compute needs has exceeded their change in compute access. This suggests academic research is likely to be increasingly constrained, compared to industry research, as compute needs increase. However, when we asked respondents how concerned they were that a lack of compute resources would be an obstacle to their AI contributions in the next decade, responses reveal little difference in concern. Figure 4 compares responses across academia and industry. Academics were slightly more likely to report being "moderately" or "extremely" concerned, but those differences are not significant. Ultimately, we find some support for a growing gap in compute access between academia and industry but no support for a higher level of concern among academics.

[4]Note that 18% of industry researchers and 9% of academics did not report compute use by either metric.

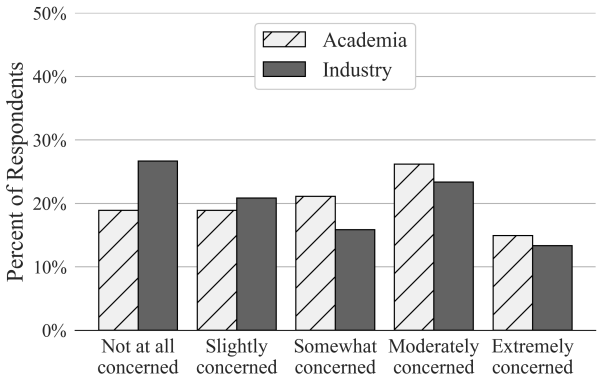

Figure 4: Respondent concern over future compute access by sector

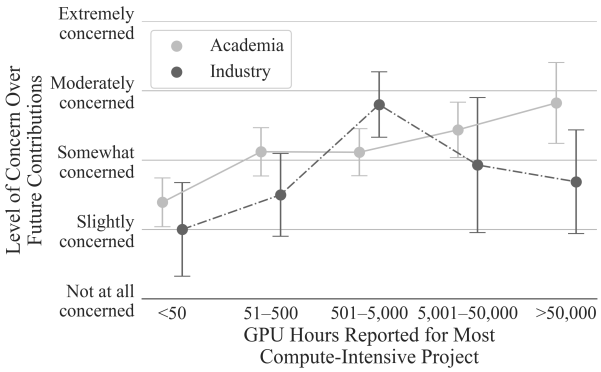

Figure 5: Mean level of concern over future compute by compute use and sector[5]

Finding 2.3. Higher Compute Use Correlates With Being More Concerned About Compute. Interesting trends emerge in the relationship between current compute use and future concerns. Figure 5 shows respondents' level of concern about insufficient compute to contribute meaningfully to future AI research against reported GPU hours for their most compute-intensive project. Respondents who used higher amounts of compute express significantly more concern about having sufficient compute to contribute to research in the future. (Differences between sectors are not significant; see appendix). Academics at the upper end of the compute use range expressed the greatest concern of any group. We also find that higher reported compute use is positively correlated with changing project plans due to lack of compute; considering compute an important factor in leaving academia; and agreeing that compute was a driver of AI progress over the last decade and will continue to be over the next decade (see appendix for details).

One explanation is that researchers' current level of compute use is influenced by self-selection: Researchers choose

[5]Note that the two lowest responses (no GPU-hours and < 50 GPU-hours) and the two highest responses (50,001-500,000 GPU-hours and more than 500,000 GPU-hours) are combined due to small sample sizes at the extreme ends of the range for GPU-hours.

to pursue work in more computationally-intensive areas or to adopt particularly computationally-intensive research methods. This self-selection could then shape levels of concern and tendencies to revise research based on compute access. But this might also mean that researchers who already use a lot of compute would be the most motivated to seek out and make use of new compute resources. In this case, attempts to provide more compute to researchers broadly could increase any existing divides between high-compute users and low-compute users.

Researchers Have a Range of Opinions About a National AI Research Resource

To increase understanding of how government could address resource gaps or inequities, and better pinpoint resource prioritization, we asked them to select among five resources they would find useful for the government to provide. Respondents most often selected grant funding (n = 325), followed by government compute resources (n = 268), then data (n = 238) and technical staff (n = 148). A possible explanation for the interest in compute, despite our finding that data and talent constrain researchers more often, is that researchers prefer flexibility when offered resources directly. Grants can be used for a wide range of purposes, while compute resources may be expended on any computation. Technical support staff could mean outside non-vetted workers, and guidelines, standards, and frameworks (n = 124) may not directly contribute to project completion.

We then invited respondents to share thoughts about the creation of national AI research resources, and received 85 responses. Most responses focused on the five resources from the preceding question, but many also offered broader suggestions about implementing government-led initiatives to support AI researchers. We engaged in focused coding and comparison methods to analyze responses. After initial evaluation, two team members independently coded responses using thematic analysis (Braun and Clarke 2006). We collated themes to build a final thematic codebook. We do not report agreement metrics as both researchers coded all responses and resolved any discrepancies.

Compute and Data Resources

Of the 85 respondents discussing government AI resources, 38 mentioned compute. A majority addressed the potential utility of compute resources to AI research: "papers are often rejected on the grounds of limited experiments, which were in fact limited because of a lack of compute." Many respondents highlighted perceived gaps between academia and industry, with one stating "[t]he amount of compute resources" in academia "is multiple orders of magnitude smaller than" at large tech firms. Multiple respondents mentioned this gap as a concern, with one also mentioning a resource gap between start-ups and large tech firms. One respondent pointed out that they have found other options: "I'd love to have compute resources, [...] but I'm intentionally choosing the projects where I'm not blocked by these things—and there are several lifetimes worth of projects in these areas."

However, eight respondents expressed doubt about pursuing a compute-intensive approach for national AI resources, suggesting compute might not actually enable that much research progress, and that "there's a ton of research that could be done on much smaller scales," or worrying it would not address existing research gaps since "not many efforts have been spent in understanding why [...] progress has been driven by computationally heavy research." Finally, five respondents were skeptical that government compute would remain accessible, supported, and cutting-edge as technology improves.

Seventeen respondents discussed data and accessibility, including the need for open, accessible datasets and related resources (e.g., source code, models, etc.) and the need for greater access to data to advance AI. Multiple respondents noted the need for diverse and accessible data sets, with one observing that "we need better privacy laws in general, but we also need much more personal human data available to researchers to make progress toward human-centered technologies." While a relatively small number of respondents specifically mentioned data issues in their answers, most of these answers emphasized difficulties accessing well-cleaned, curated, and maintained datasets.

Talent, Workforce Development, and Grants

Among 23 responses who brought up talent and workforce, many emphasized structural problems. One respondent noted that difficulties faced by foreign PhDs limit the US AI workforce: "Foreign students complete amazing work on their PhDs and then struggle to continue after graduation." Another observed that academics moving to industry creates risks of "state-of-the-art AI [becoming] monopolized and controlled by a small number of corporate entities." Several respondents were explicit that they prioritize workforce development initiatives, such as upskilling, over compute. One respondent noting "I want my tax dollars to fund 30-40 year investments in PEOPLE, not 3-5 year investments in hardware," while another explained that they needed "somebody who can help design and think through a data collection and labeling process" and "bridge the communication and knowledge gap with ML researchers."

Twenty-five respondents offered feedback on grants and their allocation. A common viewpoint was exemplified by one respondent's comment that "grant funding would be more useful than compute resources." One justification of this view was that, "government is generally bad at predicting what resources will be needed. I think it's better to give funding and let the users themselves determine what they need." Nine respondents encouraged further investment into particular areas in computer science. Some research topics included, "'small-data' algorithms that may have better utilization across the world" or "AI research that produces public goods like the prevention of catastrophes."

A few respondents noted a need for supplemental technical staff support to ensure accessibility of national AI research resources, rooted in the concern that compute resources would, while potentially providing value, have a learning and transition cost. One respondent explained, "Compute resources are not standardized enough at this

point when it comes to AI" so switching to them would require "technical staff who can manage the transition and guide development."

Guidelines, Standards, and Frameworks

Of 27 respondents who discussed guidelines, standards, or frameworks, twelve mentioned technical standards, open source tools, and evaluation. One respondent requested "testbeds that can evaluate interactive AI systems with a diverse pool of human users in realistic settings," while another wanted "specific ambitious challenges, with well defined metrics." Some respondents mentioned specific tools like "open source and pre-trained GPT3 and DALL-E." or encouraged the federal government to provide "open source code and software libraries" to help close the academia-industry research gap.

Another eight respondents discussed the need for ethical or legal frameworks for developing, using, and evaluating AI systems, often expressing a desire to see the government, "enforce standards around ETHICS in AI." Respondents noted a need for legal guidelines for AI, with one stating, "the government [...] should have an agency responsible for adapting human rules to AI." Some researchers proposed working within existing government structures to meet this need, such as "scal[ing] up NIST and help[ing] them get their message, skills, and tools out into the world."

General Suggestions

Fifty-one respondents commented on funding and provisioning, resource distribution, and suggested limitations of a government AI resource. Respondents highlighted the value of reaching out to a diverse set of researchers to support inclusive AI research. One respondent stated that not more researchers, "but more diversity of researchers could help advance the field." Multiple respondents emphasized that government-provided resources should provide "support to a wide variety of [researchers] in place of just supporting the big or known personnels." Some respondents weighed in on how the government should approach AI issues broadly, with one noting "it's important to create a task force related to finding and restricting those who come across strong AI."

These free responses offer insight into the broad diversity of perspectives from AI researchers on government provision of AI research resources. Most who provided responses welcome more government involvement in AI research, although some expressed skepticism about the government's capability to provide useful resources. While many did underscore the potential value of compute resources, others emphasized that they viewed workforce development as a higher priority, or suggested that simply scaling up compute resources will not reliably generate breakthroughs. And among those who gave specific recommendations on the implementation of government resources, the most common theme was to underscore the importance of making resources accessible and equitable.

Future Work and Limitations

Examining the compute needs of AI researchers provides a broad understanding of the field and its requirements. Our results highlight variation in compute use and prioritization; groups with unique needs or preferences include high compute users, language modelers, AI startup employees, and academics who rely solely on cloud computing. Given small sample sizes, we cannot draw strong conclusions about these groups; future work could focus on these researchers in particular, in order to better understand their requirements.

This paper focuses on disparities in compute needs between academia and industry, and across different AI subfields. Its scope is limited to research within the United States, in order to speak to salient national policy questions. While our respondents are drawn from a range of academic institutions, further work might assess differences in compute access and its implications among highly-resourced institutions and those with lower research budgets.

Finally, this paper's focus is on AI research, not AI practitioners, motivated by the impact of resource constraints on AI progress. The question of how compute access influences downstream applications of AI remains unaddressed.

Conclusions

Our survey results show that researchers do not find themselves primarily or exclusively constrained by compute access. More respondents report talent as an important factor for project success, a higher funding priority, and a limiting factor in project selection. Data availability is a more common reason for avoiding projects than insufficient compute.

There are few differences between academic and industry AI researchers in their compute use and concerns. While academic researchers report spending less money than industry researchers on compute, they report similar GPU-hour levels. Both groups report similar concern about insufficient compute allowing them to make meaningful contributions to AI research in the future. Academics are more likely to report that changes in their need for compute outpace changes in their ability to access compute, but academics are not more concerned than industry researchers that a lack of compute access will constrain their ability to contribute to future AI research. We found that heavy compute users are more likely to want additional compute than low compute users. Our respondents indicated that when it comes to government resources, they would be more receptive to compute than government-curated data or technical staff, though they would generally prefer grant funding to compute resources.

More researchers strongly agreed that compute was a major driver of the last decade of AI progress than they did other factors. But larger proportions expected most other factors we asked about to be greater drivers of the next decade of AI progress. Respondents also reported adjusting their research plans due to lack of data or talent more often than lack of compute. Certain types of AI research, most notably large "foundation models," are highly compute-intensive, and progress towards more generalizable models is presently constrained by compute. But our results suggest these issues affect a small minority of AI researchers, and for most, talent is a greater constraint than compute.

Acknowledgments

For her contributions to the design and distribution of the survey, we would like to thank Tina Huang. For their careful review, thoughtful comments, and constructive feedback, we are deeply grateful to Catherine Aiken, Autumn Toney, Drew Lohn, and many other colleagues.

References

Ahdritz, G.; Bouatta, N.; Kadyan, S.; Xia, Q.; Gerecke, W.; O'Donnell, T. J.; Berenberg, D.; Fisk, I.; Zanichelli, N.; Zhang, B.; et al. 2022. OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *bioRxiv preprint 2022.11.20.517210*.

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

Algorithmia. 2020. 2021 enterprise trends in machine learning.

Baier, L.; Jöhren, F.; and Seebacher, S. 2019. Challenges in the Deployment and Operation of Machine Learning in Practice. In *ECIS*, volume 1.

Bartoldson, B. R.; Kailkhura, B.; and Blalock, D. 2022. Compute-Efficient Deep Learning: Algorithmic Trends and Opportunities. *arXiv preprint arXiv:2210.06640*.

Benaich, N.; and Hogarth, I. 2022. State of AI Report.

Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2): 77–101.

Depp, M. 2022. Dynamic Stability: How AI will Reinforce, Not Overturn the Balance of Power. *Digital Debates*, 99.

Dey, N.; Gosal, G.; Khachane, H.; Marshall, W.; Pathria, R.; Tom, M.; Hestness, J.; et al. 2023. Cerebras-GPT: Open Compute-Optimal Language Models Trained on the Cerebras Wafer-Scale Cluster. *arXiv preprint arXiv:2304.03208*.

Dimensional Research. 2019. Artificial Intelligence and Machine Learning Projects Are Obstructed by Data Issues: Global Survey of Data Scientists, AI Experts and Stakeholders.

Ding, M.; Zheng, W.; Hong, W.; and Tang, J. 2022. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*.

Dotscience. 2019. The State of Development and Operations of AI Applications.

Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; and Tang, J. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 320–335.

Ganguli, D.; Hernandez, D.; Lovitt, L.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; Dassarma, N.; Drain, D.; Elhage, N.; et al. 2022. Predictability and surprise in large generative models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1747–1764.

Gruetzemacher, R.; Paradice, D.; and Lee, K. B. 2019. Forecasting transformative AI: An expert survey. *arXiv preprint arXiv:1901.08579*.

Hernandez, D.; and Brown, T. B. 2020. Measuring the algorithmic efficiency of neural networks. *arXiv preprint arXiv:2005.04305*.

Izsak, P.; Berchansky, M.; and Levy, O. 2021. How to train BERT with an academic budget. *arXiv preprint arXiv:2104.07705*.

Jurowetzki, R.; Hain, D.; Mateos-Garcia, J.; and Stathoulopoulos, K. 2021. The Privatization of AI Research (-ers): Causes and Potential Consequences– From university-industry interaction to public research brain-drain? *arXiv preprint arXiv:2102.01648*.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Klinger, J.; Mateos-Garcia, J.; and Stathoulopoulos, K. 2020. A narrowing of AI research? *arXiv preprint arXiv:2009.10385*.

Knight, W. 2023. OpenAI's CEO Says the Age of Giant AI Models Is Already Over. *Wired*.

McKinsey. 2022. The state of AI in 2022—and a half decade in review.

Merhi, M. I. 2023. An evaluation of the critical success factors impacting artificial intelligence implementation. *International Journal of Information Management*, 69: 102545.

Müller, V. C.; and Bostrom, N. 2016. Future progress in artificial intelligence: A survey of expert opinion. *Fundamental issues of artificial intelligence*, 555–572.

OpenAI. 2018. AI and Compute. https://openai.com/research/ai-and-compute. Accessed: 2023-03-22.

Paleyes, A.; Urma, R.-G.; and Lawrence, N. D. 2022. Challenges in deploying machine learning: a survey of case studies. *ACM Computing Surveys*, 55(6): 1–29.

Patterson, D.; Gonzalez, J.; Hölzle, U.; Le, Q.; Liang, C.; Munguia, L.-M.; Rothchild, D.; So, D. R.; Texier, M.; and Dean, J. 2022. The carbon footprint of machine learning training will plateau, then shrink. *Computer*, 55(7): 18–28.

QS Top Universities. 2023. QS World University Rankings 2023: Top global universities.

Shankar, S.; Garcia, R.; Hellerstein, J. M.; and Parameswaran, A. G. 2022. Operationalizing machine learning: An interview study. *arXiv preprint arXiv:2209.09125*.

Sharir, O.; Peleg, B.; and Shoham, Y. 2020. The cost of training NLP models: A concise overview. *arXiv preprint arXiv:2004.08900*.

Stein-Perlman, Z.; Weinstein-Raun, B.; and Grace, K. 2022. 2022 Expert Survey on Progress in AI. *AI Impacts*.

Thompson, N. C.; Ge, S.; and Manso, G. F. 2022. The importance of (exponentially more) computing power. *arXiv preprint arXiv:2206.14007*.

Thompson, N. C.; Greenewald, K.; Lee, K.; and Manso, G. F. 2020. The computational limits of deep learning. *arXiv preprint arXiv:2007.05558*.

Westenberger, J.; Schuler, K.; and Schlegel, D. 2022. Failure of AI projects: understanding the critical factors. *Procedia computer science*, 196: 69–76.

Zhang, B.; Dreksler, N.; Anderljung, M.; Kahn, L.; Giattino, C.; Dafoe, A.; and Horowitz, M. C. 2022. Forecasting AI progress: Evidence from a survey of machine learning researchers. *arXiv preprint arXiv:2206.04132*.