doi.org/10.53975/56rc-k4jt

Seeds of Science

On Scaling Academia

Jan Hendrik Kirchner[1]

Overcoming humanity's challenges will require a deep understanding of both the problem and the possible solutions. There are early indications that the scientific apparatus, which has traditionally been the primary tool for gaining deep understanding, might not be able to keep pace. In this essay, I outline a set of interventions that might help the scientific apparatus overcome existing bottlenecks, and I discuss limitations and possible implications. Centrally, I argue for systematization and automation of the research process to allow researchers to benefit from emerging technology like artificial intelligence.

Introduction

In recent years, researchers like Karnofsky (2021) and Ord (2020) make a strong case that we might be living in “the most important century”: the world is changing rapidly as new technology touches every part of our lives, and as existential risks become more and more acute. Depending on our actions today, the long-run future might look radically different. This prospect represents both a challenge to be the best version of ourselves and a promise of a potentially much better life for humanity in the future.

Whether we buy this particular argument or not, there certainly are large challenges for humanity, requiring both thought and action (Lewandowsky, 2021; Sandhu et al., 2021). Rising to these challenges will call for the use of humanity's perhaps most powerful tool, the scientific method (Saier and Trevors, 2017). Traditionally, academia, broadly construed as a community of people impartially applying the scientific method[2], is ideally positioned to use this tool. Here is Whittlestone (2018) on the potential positive impact of an academic researcher:

> A single outstanding researcher can move a field forward and make a significant contribution to solving key global problems. Beyond research, academics also have other avenues for impact, such as by influencing government policy, the priorities within their field, and the culture of society at large. (Whittlestone, 2018)

[1] PhD student, Max Planck Institute for Brain Research

[2] This definition fits both researchers working in industry and the university system.

Examples of "outstanding researchers" with a huge altruistic impact that come to mind are f.e. Norman Borlaug, whose discovery of the "miracle wheat" might have prevented the starvation of a billion people (Mann, 2019). Another example is the research team Karikó and Weissman, whose research paved the way for the mRNA vaccines that have the potential to save lives from many more things than Covid-19 (Kolata, 2021). However, in recent decades unmistakable weaknesses of the academic apparatus are starting to show (Alexander, 2014; Bloom et al., 2020; Piper, 2020; Yudkowsky, 2017), which cast doubt on the health of academia[3]. Will academia be able to rise to the challenges of the coming decades?

In this essay, I am outlining a set of interventions that might allow academia to become scalable, i.e. to produce scientific insight with constant overhead and proportional to the amount of effort/computing power invested. After observing that academia currently does not scale, I identify two factors (growing overhead, inability to exploit low-hanging fruits) that work against scalability and propose possible solutions to these problems. Finally, I investigate the feasibility of automating (portions of) academia and provide an outlook on possible changes to academic practice in the coming decades.

Inadequate science

There are some rather large research institutes: the Chinese Academy of Sciences employs 60.000 researchers, the French Centre National de la Recherche Scientifique employs 15.000 researchers, and the German Max Planck Society employs around 7.500 researchers (Nature Index, 2019). Those are not-small numbers, but they pale in comparison to the number of employees in Fortune 500 companies, which can reach into the millions (Fig. 1).

Beyond differences in the number of employees, the numbers from academia are also misleading: the 7500 researchers of the Max Planck Society are mostly just connected by some vague set of standards and values. In actuality, the society consists of a few hundred independent research groups, each with 5 to 100 researchers. And "independence" is really important here - anything like central coordination of research is perceived as limiting scientific freedom[4]. Why are 10-30 researchers the de-facto limit to how large a research group can grow? Why not thousands or tens of thousands?

Some differences between academia and the Fortune 500 are apparent immediately: Academic research requires highly trained labor and is constrained by the available funding in a given academic year. Walmart needs to provide much less training to their employees and their funding is self-generated, so we can expect scaling to be much easier here.

But there are also very strong amplification forces and accumulation of resources in academia. This is the (in)famous "Matthew effect" where the well-funded get even more funding and those with little funding get pruned out (Bol et al., 2018). Why doesn't this

[3] But see *(Guzey, 2019)* for a slightly more optimistic account of the state of the life sciences.
[4] The situation at the CAS is probably different, see f.e. the Pioneer Initiative.

dynamic "run away"? Why isn't literally everyone a Ph.D. student of the most successful scientist from their field?

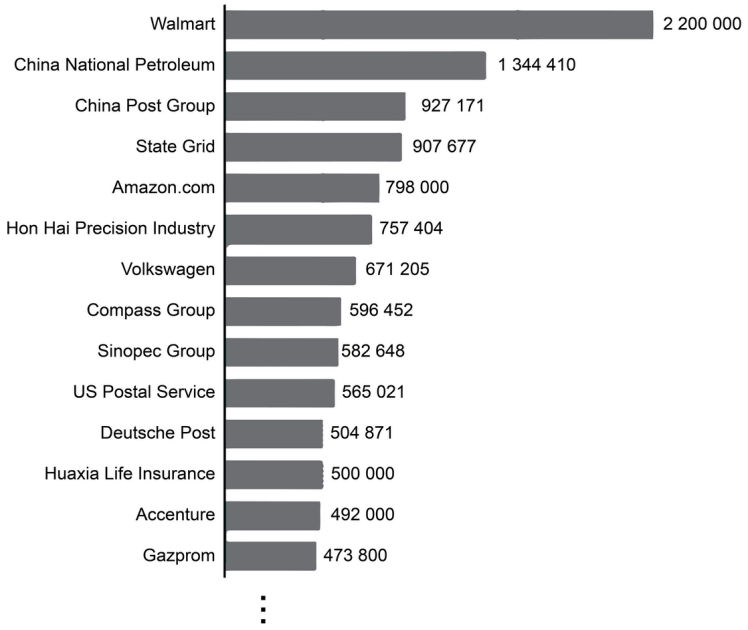

Figure 1. The number of employees in 500 Fortune companies in 2019; based on (Szmigiera, 2021).

Perhaps research questions are just not deep enough to have 1000s of people working on them? This is unlikely. Most researchers have very ambitious ideas and must content themselves to work on small portions of those questions. The Manhattan Project, the Human Genome Project, and the LHC are prominent examples of such wide-ranging projects (AMACAD, 2021; Collins et al., 2003). Indeed, it might be possible to generate deep research questions automatically (Kirchner, 2021a), so coming up with deep research questions does not appear to be the central bottleneck.

Might we conclude that we are already at some efficient frontier most of the time? Are we generating the maximal efficiency with research groups with a median size of ten people?

Increasing the group size

Let us go by the numbers first. The typical academic chain of command is very straightforward:

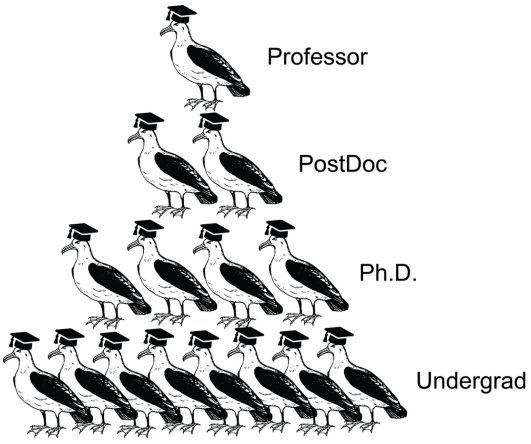

Figure 2. The academic pecking order.

When we assume that (like in Fig. 2) at each level the supervision ratio is 1:2, we can derive the average group size:

$$\text{Group size} = 2^{3+1} - 1 = 15 \quad (1)$$

Successful labs manage to scale this up by increasing the number of PostDocs or by introducing senior PostDocs/Group Leaders into the mix and by pushing how many students are supervised at each level (Fig. 3).

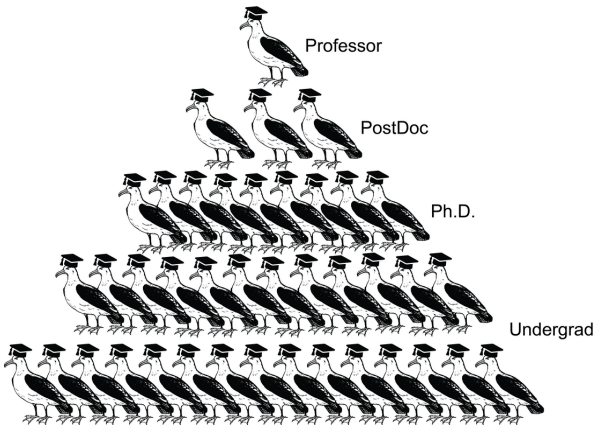

Figure 3. The pecking order intensifies.

$$\text{Group size} = 3^{3+1} - 1 = 80 \quad (2)$$

Currently, this process of extending the hierarchy does not scale. “Approximately one hundred researchers” appears to be the effective maximum of how far a traditional research group can be extended.[5] Why is that?

There are two obvious ways to increase the number of researchers in a group: **adding more levels to the hierarchy** and **increasing the number of researchers supervised at each level**. By the power of exponents, if we can add just one more level and increase the supervision ratio to 1:5, we would jump up to

$$\text{Group size} = 5^{4+1} - 1 = 3124 \quad (3)$$

Why are we not doing that?[6]

Adding more levels and increasing the supervision ratio

The natural limitation to “how deep you can make a hierarchy” is how efficiently you can propagate commands and reports up and down the chain. The natural limit to “how far you can increase the supervision ratio” is how many meetings you can squeeze into a week. Krishnan (2021) provides a great analysis of how a large organisation like the FDA, which has recently been criticized for working slow (Friedersdorf, 2021), might actually be operating at the maximum speed possible given its institutional design and overloaded responsibilities:

> Somewhere there are multiple leadership teams meeting who report to other leadership teams who report to the top, all of whom are trying to figure out what decisions should be made this week. And it's hard. Because if you have 10 projects ongoing and want to do a status update on all of them plus do review meetings, that's the week

[5] A reviewer (rightfully) requested a reference for this claim. Mryglod et al. 2016 report group sizes of U.K. biology departments from < 10 up to ~120 people and a single outlier group size of > 220.

[6] Obviously, there are more factors that influence scientific output beyond group size and a strict hierarchy is not the best way to structure things. But this objection misses the point of scaling: Having something that *does* scale, even if it looks stupid and over-simplified, beats something that doesn’t scale *(Sutton, 2019)*.

> gone. Easily. Do this whole thing 10 times in parallel, and you could easily have an organisation of 1000s but where things take forever to move. (Krishnan, 2021)

But even given that not every hierarchy can be scaled arbitrarily deep and broad, we might still be skeptical that academia is operating at the limit of what is possible. Many structural factors of academic institutional design stand in the way of systematization: Timelines are either non-existent or ridiculously poorly calibrated (Majev et al., 2021; Schoot et al., 2013). At the same time, people at the top of the hierarchy (whose time is the most valuable resource of the entire system) are overloaded with mundane tasks, paperwork, and emails (Macfarlane, 2011; Yudkowsky, 2017). Debundling workload, by getting a secretary, lab manager, or someone for illustrations is perceived as an optional luxury.

The proposed solution: The answer is simple; systematize and delegate. Create (and perfect) explicit workflows for short-term student internships, undergraduate theses, Ph.D. projects, long-running lab ambitions.[7] Have explicit timelines with regular check-ins where you see if you hit your milestones or not. Have easily accessible writing guides for papers. Set achievable, explicit goals (# of GitHub commits, # of entries in a group wiki, ...) and link career rewards to hitting these goals.

How to get there? Are we being naive? Perhaps the solution is not as easy as "just start using a project management framework from industry". Groeneveld (2020) highlights the roadblocks the author ran into when trying to implement a successful project management framework from industry in a traditional academic context. But the reasons for their failure effectively come down to the observation that traditional academia is not flexible enough to give project management frameworks a fair shot.

We should not be satisfied with that answer. Ed Boyden talks at length about his highly systematic approach to research and thinking in general (Boyden, 2007), and has consequently revolutionized neuroscience at least twice in the last two decades[8] (Boyden, 2011; Chen et al., 2015). Terence Tao, one of the most productive mathematicians alive, crowdsourced some of the work for his solution of the Erdos discrepancy problem through a massive online collaboration (Cranshaw and Kittur, 2011; Tao, 2017). And those involved with the project attribute much of the success of the Los Alamos Laboratory can be attributed to the exceptional leadership of Robert Oppenheimer (Ringer, 2007). Note how all of these examples received substantial help from outside the traditional academic system.

[7] As pointed out by a reviewer, it might be very worthwhile to *think about the design space* for organizing research groups. While hierarchies have some properties that make them beneficial, more flexible organization schemes can dominate depending on the goals.
[8] *Against all odds, as it turns out.*

Thus, systematization of the research process appears possible. We should perhaps not be surprised, as we call it the scientific method for a reason. There are great workflows and frameworks that can be applied in almost arbitrary contexts and the tools that can support researchers are becoming better and more numerous every month (Askell et al., 2021). The time appears ripe for truly ambitious scaling of research and new institutions have the potential to be the catalyzer for much of this progress (Marblestone et al., 2022).

Possibility of low-hanging fruit and lack of objective metrics

What would it mean for academia to be "efficient"? The definition from economics is usually illustrated by the joke about the economist who won't pick up the 20$ bill lying on a busy street because "if it were real, somebody else would have already picked it up". In academia, this translates to the statement "there is no low-hanging fruit": If you think you made an exciting new discovery after investing only little effort, you should be wary. Most likely, you are either wrong or somebody else already made this discovery and published it in the 1960s.

Physics is one subfield of science that appears to be "efficient" in this sense. If you think you have built a Perpetuum mobile, you are probably wrong. The same goes for mathematics, where amateur "proofs" of ancient theorems always[9] turn out to be wrong. Famously, Edmund Landau received so many incorrect amateur proofs that he prepared a form letter saying "Dear Sir or Madam: Your proof of Fermat's Last Theorem has been received. The first mistake is on page _ line _ ." so that he'd only have to fill in the page and the line (D'Souza, 2011).

But it is wrong to conclude that every subfield of science is efficient. In 1880, observing a squirrel swimming across a river gave you a good shot at getting published in Nature (Duncombe, 1881). Beyond very short journal papers (Lander and Parkin, 1966), there are also extremely short Ph.D. theses from excellent researchers (Einstein, 1906; Langerhans and Morrison, 1937; Nash, 1951). And when physics was still "young", people like Isaac Newton or Leonhard Euler single-handedly made numerous exciting discoveries.

Physics and mathematics are not young disciplines anymore, but Neuroscience, Cognitive Science, and Machine Learning are. Could it be that there is much "low-hanging fruit" that nobody is able to pick? I'm not the first to argue that academia might be inefficient in this sense. Yudkowsky (2017) argues that even a slight mismatch between "what funders value" and "what would be great for humanity" results in our inability to pick low-hanging fruit.

Beyond the general argument about potential inefficiencies, we are lacking objective metrics to evaluate what constitutes a "great scientific result". As a substitute, a lot of scientists use the proxy "time spent on a project" to evaluate the "quality of a project".[10]

[9] Exceptions like the Indian mathematician Srinivasa Ramanujan prove the rule.

[10] An important point raised by a reviewer is that with any simple metric, there is the inherent danger that it ceases to measure what we care about when put under sufficient stress (Goodhart's law).

This has been internalized to the point where some colleagues discard the straightforward and fast solution (i.e. the low-hanging fruit) in favor of a more complicated and time-intense solution. Again, I'm not the first to observe academia's unhealthy relationship with simple solutions: Machine learning researcher Andrew Ng received a surprising amount of pushback when suggesting that GPUs could be used in computer vision because this was perceived as cheating.

An exception to this pattern comes from the protein folding community, which rigorously assesses different methods for protein folding at an international conference (Kryshtafovych et al., 2021). Adopting this measure opened the door for researchers at DeepMind to develop a machine learning system that aptly overtook the previous state-of-the-art (Jumper et al., 2021). Adopting the objective metric guarantees that this great scientific result is acknowledged even when it comes from newcomers and when the results are derived in a much shorter time than the "time equals quality" proxy would suggest (Hanson, 1990).

The proposed solution: If we accept that better metrics are (part of) what is missing for scaling science, there is one prominent candidate that promises to address many of academia's shortcomings: prediction markets.

> The prediction market is a market where people can trade contracts that pay based on the outcomes of unknown future events. The market prices generated from these contracts can be understood as a kind of collective prediction among market participants. These prices are based on the individual expectations and willingness of investors to put their money on the line for those expectations. (Peters, 2021)

Adopting prediction markets in academia, as suggested already by Hanson, (1990), might solve multiple problems:

- First and foremost, we would get a computerized representation of expert knowledge/uncertainty about the state of the art in all areas of science. This opens the door to powerful quantitative reasoning about areas beyond the natural sciences (Tetlock and Gardner, 2015).
- It would provide a numerical estimate of how "difficult" a question was perceived to be before it was answered. For example, the prediction platform Metaculus asks "By 2030, will C. elegans be uploaded to the satisfaction of top computational neuroscientists?" The community converged on (through, at the time of writing, 181 predictions) a 35% probability. If a research team manages to upload C. elegans by 2025, that would indicate that an objectively difficult question was solved .

- Academics specializing in an esoteric topic currently might struggle to find a job, even though their expertise is certainly valuable (Abel et al., 2014; Morrison et al., 2011). It is just that no single company needs a monopoly on that knowledge as much as it wants to not pay an additional salary. A strong and liquid scientific prediction market would allow the Ph.D. with niche expertise to systematically beat the market on a few questions and monetize their expertise - while simultaneously increasing the power of the prediction market.

- Markets are great at predicting whether a study will replicate or not (DellaVigna et al., 2019; Menard, 2020). By leaning into prediction markets, we might be able to get out of the current dysfunctional system where the quality of a study is a rough function of the impact factor of the journal, the prestige of the group, and the visual appeal of the figures; and instead, evaluate a study based on the solidity of the results.

How to get there? The switch to a prediction market based system is a coordination problem where switching costs are incurred on those participants that move to the new system first (Farrell and Klemperer, 2007) . Most funding and prestige will be coupled to the old system for the foreseeable future, disincentivizing each individual scientist from switching.

However, not all of academia has to switch at once. While prestige and funding for picking low-hanging fruit might only follow once the larger part of academia adopts better metrics, the researchers still get to keep the low-hanging fruit regardless. If only one discipline (Machine Learning, Cognitive Science, …) or even just one sub-community (effective altruism, AI Safety) can be convinced to adopt objective metrics, they can already reap the benefits.

Setting up an academic prediction market that is technically well-done and fun to use could be highly beneficial. Perhaps Metaculus is already exactly that platform, although it uses “points” as currency which cannot serve as a source of income for experts. Alternatively, one could start a new scientific journal that has a prediction market for replication probability embedded for each paper. Again, the Metaculus Journal could serve as a blueprint. To reach a critical mass of scientific submissions, the journal could focus on publishing undergraduate or master theses in the beginning, the best of which can be of intriguingly high quality (Benito and J, 2018).

Better tools, better science

After systematizing scientific work (through collecting and refining explicit workflows) and improving metrics of scientific progress (through f.e. prediction markets) we would be in an ideal position to leverage emerging technology to further accelerate research. Consider the possibility of expanding and automating parts of the academic hierarchy (Fig. 4).

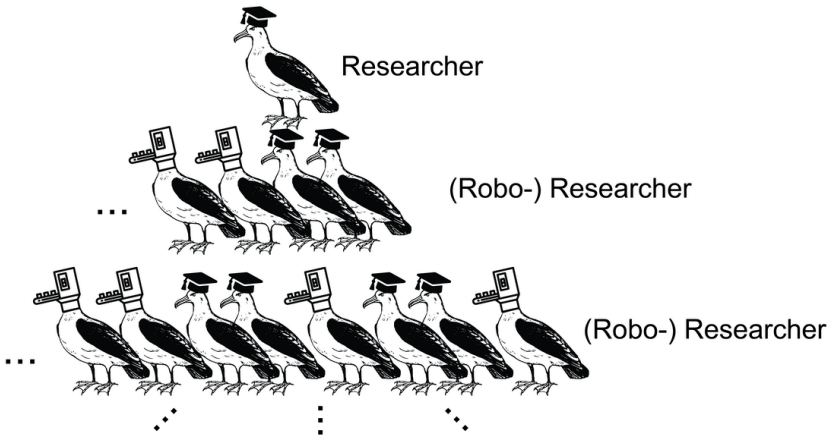

Figure 4. Automating the pecking order.

What would need to happen for this vision to become reality? Once we formalize academic workflows and establish objective metrics, a software suite composed of tools like OpenAI's Codex to automate some of the code writing (Chen et al., 2021), an automatic research assistant to help with idea generation and factorization (Kirchner, 2021b, 2021a), and Ought's Elicit for accelerating literature search, would already take us a long way towards automating substantial portions of academia.

Given the bad track record of web applications for literature search, discovery and recommendation and the comparatively low cost of academic labour (Strasser, 2021), what is the benefit of automating parts of the research process over, say, hiring all the undergrads in a given year? Three factors come to mind: consistency, flexibility, and speed. Hiring from a broad pool of academics will dilute the focus of the research, will introduce delays and lasting commitments independent of project success, and will be fundamentally limited by the speed at which humans can/want to perform routine tasks. These challenges mirror those faced by a growing startup (Barnes, 2018), but, arguably, they manifest more severely in academia where an entrepreneurial mindset is much less common (Sinclair et al., 2014).

Outlook and Conclusions

Might the academic mindset be necessary and thus make academia unfit for automation? Grace et al. (2018) performed an influential survey among machine learning experts about their timelines for when certain tasks can be accomplished by AI at the same level as humans (Fig. 5).

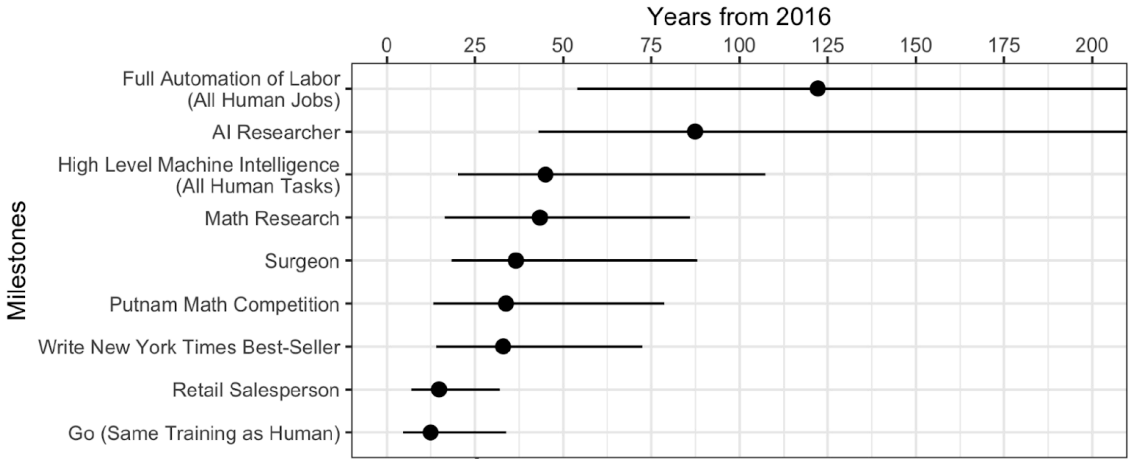

Figure 5. Timeline of Median Estimates (with 50% intervals) for AI Achieving Human Performance. Adapted from (Grace et al., 2018).

These survey results might provide a false sense of job security for academic researchers, who are estimated to be matched by automated systems only in 30 to 40 years. The estimate is contrasted by the recent progress in advanced AI, which occurred in the domain of knowledge work. AI-guided work in mathematics (Davies et al., 2021), AI-assisted coding (Chen et al., 2021), as well as extremely capable knowledge-retrieval and question answering (Borgeaud et al., 2022) all became feasible within the last 12 months. Additional reasons for why academia might feel the impact of AI before other industries are

- AI doesn't have to reach human-level for being useful and impactful in academic research.
- The amount of grunt-work in academia is underappreciated (Pacheco-Vega, 2016) and automating it away is very much in reach.
- Substantial parts of academia do not require performing experiments in the physical world but resemble software engineering.
- AI researchers do not know the typical workflows of a surgeon or of publishing an NYT bestseller. However, they do have a lot of insight into the typical workflow of a researcher.

What should we make of this? One rather obvious conclusion is that we should not expect the near future in academia to look a lot like the last ten years. We might also attempt to put ourselves in a position where we have the tools to pick low-hanging fruit as soon as it becomes feasible. Finally, we might want to be prudent and ensure that we

use future enhanced capabilities responsibly and that we are mindful of tripwires and deadlocks (Dafoe, 2021).

Gardener Comments

Mark Baum:
I think the manuscript is nicely composed and thought provoking. If the authors are interested, I think some further elaboration on why academic researchers have so far resisted the adoption of more objective project management frameworks would be valuable. I have little background in this question, but objective metrics imply the potential for unambiguously failed projects. Naturally, academics have a strong aversion to this and the academic structure accommodates failure in a number of ways. When projects arrive at ambiguous or uninteresting conclusions, they're published in little-read journals, poster-sessions, or thesis documents that few or zero people will read. How could the fear of objective failure be mitigated? Maybe there are lessons from the technology industry.

Anonymous1:
I'm of two minds on this manuscript. I don't think the answer to the problems in science is more scale enlargement, more industry mimicking, more market forces, and more hierarchy (supervision ratio). While I speak mainly for the disciplines I know (psychology, cognitive science and philosophy), I think it also partly applies to the more mature or applied sciences. Scientists have to be able to move swiftly and try out different things, without needing to check with an elaborate pecking order, or a larger group, or comply with project management requirements. They also have to be able to collaborate as *independently* as possible instead of in hierarchies, because they need to be able to criticize each other fundamentally without fear of repercussions (caused by dependency/power relations). The ratio PhDs vs PIs is already dangerously high, causing most science to be done by (very) early career researchers with too little guidance. PIs take on too many PhD students, while postdocs often don't have an incentive to provide thorough support and guidance, because they have to make sure they produce (first or last author) papers themselves (on PhD projects, those author positions are often already reserved for PhD students and the PI). This piece ("Academic Precarity and the Single PI Model") by Romain Brette is a must-read on these issues. The system that the authors propose does nothing to address these challenges of current academic science. I'm against the push to formalize, systematize, streamline, scale, and just 'project-manage' science. Some domains of science may be of this run-of-the-mill, paradigmatic kind, that could benefit from automatization and project management. But most of (the front line of) science is (and has to be) messy and inefficient. It needs to allow for randomness (cf. Roger's Bacon et al in SoS) and dead ends. Its goals emerge while doing it, instead of being planned beforehand. Setting clear goals beforehand is what funders want, and doubling down on that mindset (as the authors propose) will not lead to more research on "what would be great for humanity".

The thing I did like about this paper is the authors' proposal to use prediction markets as potential new metrics in science. I'd gladly read and support a full paper on how this could be worked out, as well as on how the machine learning tools ("robo-researchers") mentioned at the end (eg Elicit) will impact science. I'm very enthusiastic about what these tools can mean for science, but don't share the problem statement and solution in the first half (and title) of this manuscript.

Dan James:
I found this an interesting discussion/proposal and an important topic. Undoubtedly an idealized version of the scientific method is our best tool for gaining knowledge, but in practice the scientific method itself is not without its limitations - in particular at the level of hypothesis generation and a potential lack of importance placed on observations that lie outside the main research programme or paradigm (Castillo, 2013).

Perhaps a mention of this would be a relevant inclusion in the paper? A vivid contemporary example of how this problem (of an unwillingness to pursue/support alternative hypotheses) can have potentially disastrous effects on knowledge is the MRNA research, actually mentioned in the paper, of Katalin Karikó, which was for over 10 years seen as a dead end by other more senior researchers, to the extent that she had to accept demotion and a pay cut to even stay within the research community.

Its not clear to me that scaling academia is a necessary condition toward improving scientific research, even less that it is a sufficient condition. Scaling the current academic structure, which as this paper points out is highly hierarchical, seems to run the unwelcome risk of scaling the problems by potentially prioritising quantity and increased outputs of enlarged research teams over quality and meaningful research that genuinely expands knowledge. If the only metric used is journal publication there is a risk of ‘Zombie Science’ (Berenbaum 2021) This paper seems to accept an academic hierarchical structure without question, but could alternative models better encourage teamwork/research? I’m thinking here specifically of an Authority Gradient analysis of teams that has proved very effective in other sectors such as Medicine and Air Safety.

The other candidate mentioned (for scaling academia) is prediction markets, which I found to be a welcome and unexpected surprise in a paper that I thought was going to be mainly about improving efficiencies and scale in a traditional hierarchical system.

What was not made clear was whether the use of academic/scientific predictions markets should be open to all or just to those already involved in academic research. If not open to all, why not? Also, whilst prediction markets are appealing for many of the reasons stated in the paper, their flaws should be pointed out, especially their susceptibility to ‘gaming’ (Y Chen, 2009)

Anonymous2:
This paper presents some very intriguing ideas and is well worth publishing. In particular, the concept of involving prediction markets in the academic process shows great promise. Page 8, bottom, seems to have a link missing, and the last paragraph of page 9 should read “platform” for “plastform”.

Phil Wilson:
There is an interesting idea somewhere in the paper, but before I could recommend publication, the idea needs to be better explained and supported with better examples.

- the section "inadequate science" is confusing, because it seems to first equate "employees" and "honorary members" as well as "corporations" and "academic societies", before backtracking on that comparison. If it's just about numbers, simply say so.

- the sentence before equation (1), and the equation itself: this "analysis" is both begging the question and cherry picking the data. Cherry picking, because what about groups with more than one professor? or with no undergrads? Begging the question, because of course you get about the right answer if you assume a nice supervision ratio. What is this assumption based on? It surely varies hugely from field to field. It seems like it was chosen to give the "right" answer of 15.

- the "proposed solution" at the bottom of p.6: What about the wealth of evidence, both from data and from personal recollection of successful scientists, that such number-chasing is precisely opposed to creative scientific output? You will end up reifying these pointless little targets, and forget the important goal of doing good science.

- the sentence about Terence Tao on p.7 does not support the point at all.

- the sentence "note how all of these examples received substantial help from outside the traditional academic system" on p.7: No they don't. The people helping Tao with his crowdsourcing were almost exclusively university academics. Why does Oppenheimer count as "substantial help from outside the traditional academic system"?

These issues can be addressed, and then I think a good paper will result.

Ted Wade:
This paper is original, clear, and thoughtful, with promising suggestions and a good reference list for how to improve scientific culture. I can’t speak to the novelty of its ideas separately, but it seems possible that their combination under the theme of scalability might be.

One of the paper's topics, prediction markets, sometimes seems like a panacea (like blockchains!) for the rationalist community, as the gadget that can solve many kinds of problems. There are interesting issues in how such markets might help science. Consider their use, as the paper suggests, for evaluation metrics. Prediction markets are only as good as their questions, which are only as good as timely verifiability of their resolution criteria. Arguably, whether a scientific issue is resolved or not is a matter of predominant opinion, and often opposition opinions never go away (look at current climate science).

In that case, a prediction resolution criterion might have to say something like: on <date> at least <X>% of <domain practitioners> agree that <some assertion is true>. Even if we can specify who the domain practitioners legitimately are, this does not get us away from a possible bandwagon effect.

Perhaps we should focus scientific efforts on posing questions that do have a solid resolution criterion. In essence, we are already trying to do that in every grant application and research plan, but it's very hard and a realistic attitude leads one to waffle.

Investigators have an incentive to provide resolution criteria for their research questions, and a new trend is to do this with pre-registration of hypotheses. Generally they predict the outcome of individual experiments. The truth value of a broader hypothesis, or of a higher-level theory is much harder to reduce to a simple criterion. Also, a lot of research is exploratory, working down a branching tree of possibilities, with no crisp resolution criteria available. Is exploratory research valuable? Very much. It can lead to breakthroughs and is a big source for new ideas. Exploration can also devolve into p-hacking.

In a better science workflow, then, prediction market evaluation might be more useful for more mature lines of inquiry. Another workflow issue is the timing of when a bet on a prediction gets resolved. This can't happen before the research is executed, so it can't be used for funding decisions. But you could make grant peer reviewers bet. If they refuse to bet on a proposal, then they are not referees for it.

The paper mentions better employment for niche experts as a side effect. Going beyond that, being a general contrarian scientist-predictor could be a career if the prediction markets pay in money. That might be a good thing. Gadflies are useful.

If markets pay only in prestige, might that lead to runaway concentration of power like the current prestige-dominant system?

Academic science workloads are notoriously high. Adding a duty to evaluate and bet on prediction markets would just make workload worse. How would that play out with

different payment schemes? Would it fail if treated as unpaid labor, like peer review is now?

We know that the public generally misunderstands, and all too often mistrusts, the scientific process. Would that improve if we were seen betting on outcomes? Language to describe the practice would need to avoid implications of gambling, betting, or (!) corruption.

Anonymous3:
While I appreciate the discussion of the relative ease and difficulty with which new strategies could be applied to academia, I see two primary areas where this could be improved as well as maintaining reservations on the 'political' implications. 1 The article presents a description of academia without inclusion of non-science disciplines. This is an issue because a large portion of academics are not scientists and therefore would not directly benefit as well as ignoring the possible issues present there. As the author describes the difficulty of applying AI techniques to novel writing, there should be similar issues with producing literary criticism, historical analysis, and music production. 2 The author makes a number of comparisons to economic strategies, going as far as supporting market based mechanisms for determining the difficulty of problems. Similar to modern crises surrounding rampant scaling and construction of hierarchies outside of science, there does not seem to be an acknowledgement of the value that scaling might have. Outside of certain huge projects, such as Human Genome, Manhattan etc, the value of scaling could only be highly systems thinking areas such as neuroscience and developmental biology. This is followed by its own issue of a potential loss of perspective as certain theoretical frameworks are taken on by entire institutions limiting possible understanding (see theory-limited thinking in philosophy of science). My personal reservations are partially derived from point 2 as these strategies are reliant upon and reinforce hierarchy within academia. This structuring of knowledge and people can be harmful for the individuals as well as the interactions between them, allowing for unfair power dynamics and control over the direction of projects.

Michael M. Kazanjian:
The author ought to address the problem of credentialism as an obstacle to progress. An MBA who taught in a university college of commerce, did great research and submitted a physics paper to a physics journal. One or more colleagues who taught in the university physics' department, found the paper fascinating. He submitted it, and the journal liked but rejected it because MBAs write about business, not physics. Apparently, physicists will not read a physicist-refereed paper by an MBA. Credentialism must stop. Abraham Maslow lamented that carpenters' unions say that only carpenters may touch wood, and they must touch only wood. FBI Director J. Edgar Hoover wanted agents to be diversified with many talents. The CIA has agents and officials qualified to teach in most colleges.

Article covers lots of ground in very short space. It's quite comprehensive, but that may be good. I wonder if it is generally too quantitative? Might the math be more verbalized. Not being a mathematician, I cannot answer that question.

Beriukay:
On p6, "Set achievable, explicit goals", it seems like this section might want to address how scaling might run afoul of optimizing on such easily-measured goals that lead it astray from the actual goal of scaling research. On p10, "However, not all of academia has to switch at once." One possible avenue is to just make prediction markets the arbiter of disputes between prominent scientists. We have already seen interesting youtube results along this line, with the Chain Fountain Dispute between Steve Mould and Mehdi Sadaghdar.

Rohit:
I think the article is interesting, but could use more fleshing out of the solution space. As it stands it reads as the application of project management tools + prediction markets to help make academia more efficient. But a question is whether that's the actual problem in the first place. For instance, I looked at the costs of coordination (Part II of https://www.strangeloopcanon.com/p/hierarchical-growth-trade-offs) which has a look at the costs of hierarchical vs decentralised coordination.

To truly address the question I'd say also requires a view on whether the questions being asked of academia needs larger teams now, and what success stories here might look like - e.g. the hadron collider and search for higgs boson.

Anonymous4:
This paper presents some very intriguing ideas and is well worth publishing. In particular, the concept of involving prediction markets in the academic process shows great promise.

Partha Ghosh:
The necessity and priority of scaling (compared to, say, diversity of background, or defining incentives/ success criteria) are not established. Using project management tools and automation/ AI for efficiency are indeed sensible, but hardly novel.

Mark:
I like the direction this article is taking, but I feel it would benefit a lot from better justifying its core thesis (that scaling academia is the right notion to solve the problems science faces today) and, by orienting less around high level concepts like scale and group management, and instead dealing more into the foundational issues of science that make improving it a challenge. For example, how we can design scientific endeavors to learn more things from them, and how generalized systematization of knowledge can be done practically.

Acknowledgments

I would like to thank the reviewers and the editor for their thoughtful comments and feedback on this article.

References

- Abel, J.R., Deitz, R., and Su, Y. (2014). Are Recent College Graduates Finding Good Jobs? (Rochester, NY: Social Science Research Network).
- Alexander, S. (2014). The Control Group Is Out Of Control.
- AMACAD (2021). Exploring the Future of International Large-Scale Science.
- Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., et al. (2021). A General Language Assistant as a Laboratory for Alignment. ArXiv211200861 Cs.
- Barnes, J. (2018). Council Post: The Challenges Faced By Fast-Growing Companies, And How To Solve Them.
- Benito, M., and J, O. (2018). Methodological Quality and Characteristics of the Undergraduate Psychology Theses of a Private University of Peru. J. Educ. Psychol. - Propos. Represent. 6, 321–338.
- Bloom, N., Jones, C.I., Van Reenen, J., and Webb, M. (2020). Are Ideas Getting Harder to Find? Am. Econ. Rev. 110, 1104–1144.
- Bol, T., Vaan, M. de, and Rijt, A. van de (2018). The Matthew effect in science funding. Proc. Natl. Acad. Sci. 115, 4887–4890.
- Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Driessche, G. van den, Lespiau, J.-B., Damoc, B., Clark, A., et al. (2022). Improving language models by retrieving from trillions of tokens. ArXiv211204426 Cs.
- Boyden, E. (2007). How to Think.
- Boyden, E.S. (2011). A history of optogenetics: the development of tools for controlling brain circuits with light. F1000 Biol. Rep. 3, 11.
- Chen, F., Tillberg, P.W., and Boyden, E.S. (2015). Expansion microscopy. Science.
- Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. (2021). Evaluating Large Language Models Trained on Code. ArXiv210703374 Cs.
- Collins, F.S., Morgan, M., and Patrinos, A. (2003). The Human Genome Project: Lessons from Large-Scale Biology. Science.
- Cranshaw, J., and Kittur, A. (2011). The polymath project: lessons from a successful online collaboration in mathematics. CHI.
- Dafoe, A. (2021). AI Governance: A Research Agenda | GovAI.
- Davies, A., Veličković, P., Buesing, L., Blackwell, S., Zheng, D., Tomašev, N., Tanburn, R., Battaglia, P., Blundell, C., Juhász, A., et al. (2021). Advancing mathematics by guiding human intuition with AI. Nature 600, 70–74.
- DellaVigna, S., Pope, D., and Vivalt, E. (2019). Predict science to improve science. Science.
- D'Souza, D. (2011). Margins of a theorem.

- Duncombe, C. (1881). Squirrels Crossing Water. Nature 23, 485–485.
- Einstein, A. (1906). Eine neue Bestimmung der Moleküldimensionen. Ann. Phys. 324, 289–306.
- Farrell, J., and Klemperer, P. (2007). Chapter 31 Coordination and Lock-In: Competition with Switching Costs and Network Effects. In Handbook of Industrial Organization, M. Armstrong, and R. Porter, eds. (Elsevier), pp. 1967–2072.
- Friedersdorf, C. (2021). The Death Toll of Delay.
- Grace, K., Salvatier, J., Dafoe, A., Zhang, B., and Evans, O. (2018). When Will AI Exceed Human Performance? Evidence from AI Experts. ArXiv170508807 Cs.
- Groeneveld, W. (2020). Five reasons why agile and academia don't go together.
- Guzey, A. (2019). How Life Sciences Actually Work: Findings of a Year-Long Investigation. Guzey.Com.
- Hanson, R. (1990). Could Gambling Save Science?
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589.
- Karnofsky, H. (2021). The "most important century" blog post series.
- Kirchner, J.H. (2021a). On Automatic Ideas - by Jan Hendrik Kirchner.
- Kirchner, J.H. (2021b). Making of #IAN.
- Kolata, G. (2021). Kati Kariko Helped Shield the World From the Coronavirus. N. Y. Times.
- Krishnan, R. (2021). Meditations On Regulations: Quis Custodiet Ipsos Custodes.
- Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. (2021). Critical assessment of methods of protein structure prediction (CASP)—Round XIV. Proteins Struct. Funct. Bioinforma. 89, 1607–1617.
- Lander, L.J., and Parkin, T.R. (1966). Counterexample to Euler's conjecture on sums of like powers. Bull Amer Math Soc 72, 1079.
- Langerhans, P., and Morrison, H. (1937). CONTRIBUTIONS TO THE MICROSCOPIC ANATOMY OF THE PANCREAS. Bull. Inst. Hist. Med. 5, 259–297.
- Lewandowsky, S. (2021). Climate Change Disinformation and How to Combat It. Annu. Rev. Public Health 42, 1–21.
- Macfarlane, B. (2011). The Morphing of Academic Practice: Unbundling and the Rise of the Para-academic. High. Educ. Q. 65, 59–73.
- Majev, P.-G., Vieira, R.M., Carollo, A., Liu, H., Stutz, D., Fahrenwaldt, A., Drummond, N., and Group 2020/2021, M.P.P. survey (2021). PhDnet Report 2020.
- Mann, C. (2019). The Wizard and the Prophet by Charles Mann: 9780345802842 | PenguinRandomHouse.com: Books.
- Marblestone, A., Gamick, A., Kalil, T., Martin, C., Cvitkovic, M., and Rodriques, S.G. (2022). Unblock research bottlenecks with non-profit start-ups. Nature 601, 188–190.
- Menard, A. de (2020). What's Wrong with Social Science and How to Fix It: Reflections After Reading 2578 Papers. Fantast. Anachronism.
- Morrison, E., Rudd, E., and Nerad, M. (2011). Early careers of recent U.S. Social

Science PhDs. Learn. Teach. 4, 6–29.
- Nash, J. (1951). Non-cooperative games. Ann. Math. 286–295.
- Nature Index (2019). The top 10 research institutions for 2018.
- Ord, T. (2020). The Precipice.
- Pacheco-Vega, R. (2016). On doing the grunt work in academia.
- Peters, K. (2021). Prediction Market Definition.
- Piper, K. (2020). Science has been in a “replication crisis” for a decade. Have we learned anything? - Vox.
- Ringer, R.C. (2007). Lessons in Leadership: Robert Oppenheimer and the Los Alamos Laboratory. Organ. Manag. J. 4, 25–42.
- Saier, M.H., and Trevors, J.T. (2017). Science, Innovation and the Future of Humanity. J. Mol. Microbiol. Biotechnol. 27, 128–132.
- Sandhu, H.S., Arora, A., Sarker, S.I., Shah, B., Sivendra, A., Winsor, E.S., and Luthra, A. (2021). Pandemic prevention and unsustainable animal-based consumption. Bull. World Health Organ. 99, 603–605.
- Schoot, R. van de, Yerkes, M.A., Mouw, J.M., and Sonneveld, H. (2013). What Took Them So Long? Explaining PhD Delays among Doctoral Candidates. PLOS ONE 8, e68839.
- Sinclair, J., Cuthbert, D., and Barnacle, R. (2014). The entrepreneurial subjectivity of successful researchers. High. Educ. Res. Dev. 33, 1007–1019.
- Strasser, M. (2021). The Business of Extracting Knowledge from Academic Publications.
- Sutton, R. (2019). The Bitter Lesson.
- Szmigiera, M. (2021). Number of employees: largest companies in the world.
- Tao, T. (2017). The Erdos discrepancy problem. ArXiv150905363 Math.
- Tetlock, P.E., and Gardner, D. (2015). Superforecasting: The Art and Science of Prediction by Philip E. Tetlock.
- Whittlestone, J. (2018). Considering becoming an academic? Read this first.
- Yudkowsky, E. (2017). Inadequate equilibria: Where and how civilizations get stuck (Machine Intelligence Research Institute).