Autonomous researchers will strain peer review, funding, and oversight

Jason Fantl
Averroes Labs

October 2026

Summary

Autonomous AI researchers could greatly accelerate AI safety research, but even if we solve sabotage, they will strain the research ecosystem. They will multiply the volume of research and make researchers homogeneous. If nothing changes, peer review will collapse, false positives will rise, funding will concentrate, and humans will lose track of the science. We should start valuing replication, preserving diversity, and anchoring research to human understanding now. I think this needs a dedicated metascience organization.

Originally published on LessWrong on 21 October 2026.

Frontier labs are aiming to build autonomous researchers, and they are making progress. OpenAI announced in September 2026 that they had built a "research intern", and their stated goal is a fully automated AI researcher by March 2028. Independent forecasters put full automation of AI R&D a few years later, around 2031 to 2032. Autonomous researchers will amplify many issues that already exist in science and introduce new issues. We already see AI having an impact on research and publishing, primarily automating many steps of the research process, increasing the volume of publications, and weakening the peer review process.

We should think more deeply about what autonomous researchers will look like and what they mean for the ecosystem of AI safety research. While there is huge potential to accelerate progress, there is also the possibility that existing infrastructure collapses under this new weight.

The failures we consider here are relevant even with controlled autonomous researchers, but I highly recommend looking at AI control, which addresses sabotage, a major roadblock to automated AI safety research. This is covered well by Carlsmith and Benton et al.. Everything below is a concern even if AI is not trying to sabotage.

We should also keep in mind that autonomous researchers may cause the ecosystem to reorganize in ways we can't predict, for better or worse. I think it is important to keep our minds open as AI enables previously unimaginable technology and coordination. Below are some wild guesses to get the imagination flowing, but the rest of this paper focuses only on the potential failures of today's research ecosystem.

Some interesting ideas for new coordination mechanisms.

1. Publications shared at the size of individual experiments instead of papers. Atomized research would speed up collaboration and collective exploration. We see a form of this taking place in math, with Lean proofs of theorems posted without an associated paper, and then others immediately building on those proofs.
2. Research projects worked on by millions of transient collaborators, with authorship assigned to the lucky agents that made contributions along the critical path to the eventual publication. We again see

an early version of this in math: the Equational Theories Project settled over 22 million implications between algebraic laws, with results contributed by many people and verified in Lean.

3. Living citations between all related work. As research is refined or falsified, the papers that cite it would also update, potentially in a cascade. This would create a living network of experiments and claims, collectively updating to new information and representing the clearest snapshot of the world that science has produced up to this point.
4. Papers are staked with a dollar amount (or some "reputation" points) roughly representing the authors' confidence and effort, and anyone who falsifies the paper is awarded the stake, creating a market for true statements. This keeps out low-effort work by imposing a cost on submission and incentivizes replications through the potential reward from falsification. However, this only moves the verification problem to the falsification step, so it would need to be combined with another mechanism. Related ideas are prediction markets on replication and the ERROR project, which pays bounties for errors found in published papers.
5. A user pays to have their paper reviewed (small payments for AI review, larger amounts for human review), and the more favorable the review, the more visibility the paper gets on a forum. This, along with a few other ideas, was proposed by @Raemon in An Alignment Forum for AIs? (or: Verification in the Age of Slop).

Today's infrastructure

Today's AI safety research infrastructure is largely the same as the rest of academia's: researchers from all over (university labs, non-profits, industry, think tanks, independents) receive funding to write papers, which they then submit to journals and conferences in the hope of publication, and then repeat the process. The publication platforms attempt to filter for high-quality work, first by desk-rejecting papers that fail to meet minimum standards, and then through peer review, which ML conferences increasingly staff by requiring submitting authors to serve as reviewers.

It is also standard in ML to upload work to preprint servers such as arXiv. This lets researchers claim ideas that might otherwise get scooped during the long publication process, and share their research before it has been peer reviewed. The AI safety community has also developed the helpful habit of writing a blog post alongside each paper; the paper has the same content but is easier to digest.

Research dissemination is now done through a combination of Twitter, blog posts, arXiv, Google search, and, increasingly, LLM literature reviews. Journals and conferences play a smaller role in dissemination because their timelines are too long for AI safety research. The publication process can take several months to a year, but current AI safety research moves on the order of a few months, and this is only getting shorter.

Autonomous researchers

If we expect autonomous researchers to improve and take over more of the research (as we are seeing), what qualities can we expect to see across science?

- **More research**. The number of researchers will scale with the number of GPUs, and the researchers will work non-stop. Publication rates will be unprecedented.
- **Homogeneity**. There is a limited number of SoTA models that human researchers choose to use, so the potentially hundreds of thousands of autonomous researchers will be copies of the same limited number of models.
- **Quality**. When autonomous researchers first begin to do most of the work, they will produce lower-quality work (the threshold to replace humans is below human level), but eventually they will produce

higher-quality work. How much they will improve, and how quickly, is an important unknown.

Because research quality is uncertain, I think it is best to assume that research will be submitted in large volumes at every level of quality. This includes everything from crackpot papers to paradigm-shifting papers. I assume the same for peer reviews.

Autonomous researchers will also allow people with any background to do research in any field, although subject-matter experts will have an advantage because they can guide their autonomous researchers' thinking. This advantage will shrink as the AIs become more capable and disappear completely once AIs can match the experts' research quality. This means we should expect research, including high-quality research, to come from everywhere.

The expected issues

If we continue to use the existing scientific ecosystem without changes, and autonomous researchers become widespread, here are the issues that I suspect will be the most pressing.

- **Peer review will collapse**.
  - **Not enough human experts**. The volume of submissions will increase dramatically, while the number of expert reviewers will stay roughly constant, forcing the use of AI reviewers. Venues are already capping submissions per author.
  - **Correlated reviews**. If we use a small number of models to review papers (which may also have been written by those same models), the homogeneous reviewers will have correlated blind spots. The number of reviews would no longer reflect how many perspectives an experiment has been subject to. This may be a short-term problem, as peer-review does not require diversity of thought and eventually a capable model will be able to provide a single strong review. Replication by the same models still helps, since it catches noise and sensitivity to parameter choices, but the replicator shares the original's design and analysis decisions, so it inherits the same blind spots.
- **Exaggerated claims and false positives will increase**.
  - **File Drawer Problem**. AI can test many more hypotheses, methodologies, and combinations of data subsets, which greatly amplifies the Garden of Forking Paths (each data-informed decision increases the chance of producing a significant result) and the File Drawer Problem (not publishing null results inflates reported effect sizes and raises the chance of publishing a false positive). This will be true for the individual researcher and for the field. The individual will test more hypotheses, and the field will have more researchers testing the same hypotheses.
  - **Undetected errors**. There are a number of ways in which errors introduced in AI-written papers could go undetected. These include errors concentrated where reviewers are least likely to catch them (because training optimizes against reviewer judgment), and entirely alien mistakes that reviewers have never seen. These are a few examples drawn from *Automated alignment is harder than you think*, which lists more.
  - **Reward hacking**. AIs have a tendency to reward-hack, and this has stayed true even as their capabilities have improved. Autonomous researchers may re-run experiments, remove outliers, or choose the most favorable statistical method to improve their results.
- **Journals and conferences will be pointless**.
  - **Paper half-life**. Publications will become outdated faster and faster, until the time it takes to publish in a journal or conference makes publishing there pointless.

- **Proxies for researcher quality will become unreliable**.
  - **Publication and citation count.** The cost to publish a paper and create citations will drop to almost nothing, so publication and citation counts will become increasingly unreliable. This proxy was already weak. The bibliometrics community's own Leiden Manifesto warns against judging researchers by citation counts, and citations can be bought on Google Scholar.
  - **Reputation.** Initially, reputation will be increasingly used to allocate funding, since it is one of the few signals of expertise that predates AI. Existing status bias will be amplified. But as models become more capable and benefit less from an expert advisor, human reputation will become less reliable. Reputation will still be useful, but it will shift from humans to models.
- **Funding will concentrate**.
  - **Funding concentrates to top researchers (short term)**. While autonomous researchers are still acting as research interns and human input can still make a difference, the top human researchers will receive most of the funding. Their output will grow as they delegate more work to AI, and we will want to make sure they are not bottlenecked by funding, so more money will go to them. New researchers will still get funding, but less than they could use, since that money is better spent on established top researchers.
  - **Funding concentrates in frontier labs**. Grant givers can specify the research direction and success criteria, then directly fund autonomous researchers. A few experts might be employed to decide that direction, but eventually autonomous researchers will be able to pick strong research directions themselves, and grant givers will simply pour money into the black box of "science" and get useful results out. Eventually, no human researchers would be needed at any step of the process, and all the funding would go to the model providers. It is unclear how big an issue this is. It hands more control to the labs whose systems we are trying to oversee, though the concern is smaller if labs keep doing and publishing safety research, as they have so far.
  - **Fewer research directions.** If the same few models are all looking at the same published research, they will converge on the same "most promising" research direction. The fewer distinct models we have and the more autonomous researchers we fund, the more funding will be concentrated in the same few research directions.
- **Humans will stop understanding science**.
  - **Scalable Oversight**. In the same way that a layperson cannot understand a niche field of science today, the whole of humanity will not be able to understand any field of science in the future. We will need to either gate the progress of science on human understanding, let it accelerate even as we become unable to comprehend its results, or find some way to enforce legibility without slowing down progress.
  - **Gradual Disempowerment**. If AI writes the papers and only other AIs read them, then humans are not in the loop at all. Science could drift from human control. Without participating, we will not understand the field well enough to decide where research should be focused, so we will hand that control to AI.

Takeaways

Autonomous researchers have huge potential to speed up progress in AI safety, but we should consider the issues they might introduce as well. Each issue above assumes no intervention. I suspect some have simple fixes, while others will require more thinking.

The biggest issue I see is that the volume of research produced by autonomous researchers will disrupt much of our existing infrastructure. This will be compounded by model homogeneity, which will introduce issues science has not dealt with before. New results will vastly outpace our ability to verify them. We should be proactive in valuing replication and review now, before the increase in volume forces a rushed response. Diversity of thought will shrink. We should be developing diversifying mechanisms before this issue silently slows science. Humans will be distanced from science. We should find automated ways to anchor research to human understanding so we can let autonomous researchers move as fast as possible without loss of oversight.

Most of these issues come from running a human-scale system at machine scale, so I don't think modifying existing venues will be enough. The Alignment Journal is a good step, paying reviewers and incentivizing shorter publication timelines. But it still relies on human reviewers for the final verdict, which is what I expect to break first. It also doesn't address issues like the file drawer problem or human oversight.

I think we need a metascience organization for AI safety research, one that experiments with how research is published, reviewed, and funded, and then builds the infrastructure for AI-driven safety research.

If you are working on this, want to, or think something here is wrong, I'd love to hear from you. I'm especially interested in how we might address these issues, and whether any are already being addressed.

Acknowledgements

Many thanks to Logan Riggs, Austin Morrissey, and Ethan Fantl for feedback on these ideas.

Thanks also to the Cambridge Boston Alignment Initiative (CBAI) for providing funding, guidance, and office space for work and community.

Edit (Oct 29): Added the sentence on how Correlated reviews are likely a short-term problem. Thanks to @Jess-Riedel for his helpful comments that prompted this edit.