Agents let AI safety share experiments hourly, not just papers monthly

Jason Fantl
Averroes Labs

September 2026

Summary

Today, AI safety research is shared primarily at the scale of papers, creating collective feedback loops that take weeks or months. I propose an agent-based research approach that also shares progress at the scale of individual experiments, allowing agents and researchers to continuously replicate, extend, critique, and build upon one another's work. By increasing the granularity of collaboration, we can potentially reduce the collective research feedback loop to hours. We can start this today. The long-term vision is to turn the entire AI safety community into a single, continuously coordinated laboratory whose research accelerates as the community grows and agents become more capable.

Originally published on LessWrong on 17 September 2026.

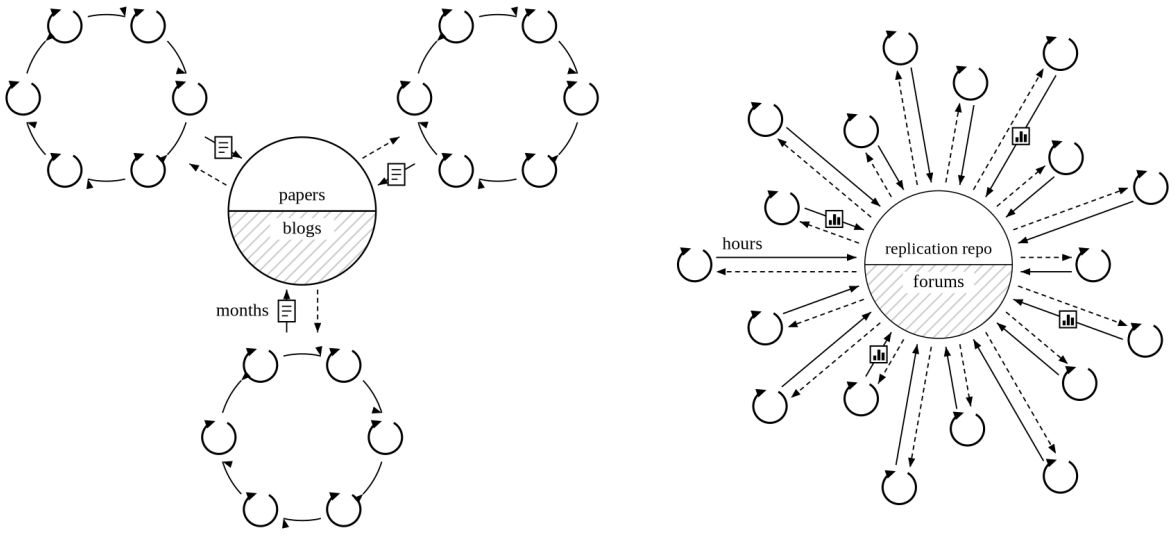
I assume the nature of research will dramatically transform as autoresearchers and swarms begin to match the quality of existing researchers. We need to consider how to prepare the field for this change, and this is one such proposal. I have not trial-run any of this yet and would like feedback before I do.

The idea

People spend months creating a publication (blog, paper), which is then shared on LessWrong or arXiv, sometimes at a conference. People comment on the blog post, tweets are made about recent papers, conversations are had in Slack channels, and the field moves forward. As models improve, research will speed up and publications will come out faster, shrinking this research loop from months to weeks.

I think we can do more. Bigger models speed up individual agents, but we should also speed up the collaboration loop. Instead of restricting agents to work on a paper for weeks or months at a time before sharing, we should allow all agents and humans to collaborate continuously. We atomize the progress into smaller units so the feedback is faster and more reactive. Instead of papers published every month, it's individual experiments every hour.

In spirit I want to turn the entire AI safety research community into one large lab, replacing the disjoint labs we have now.



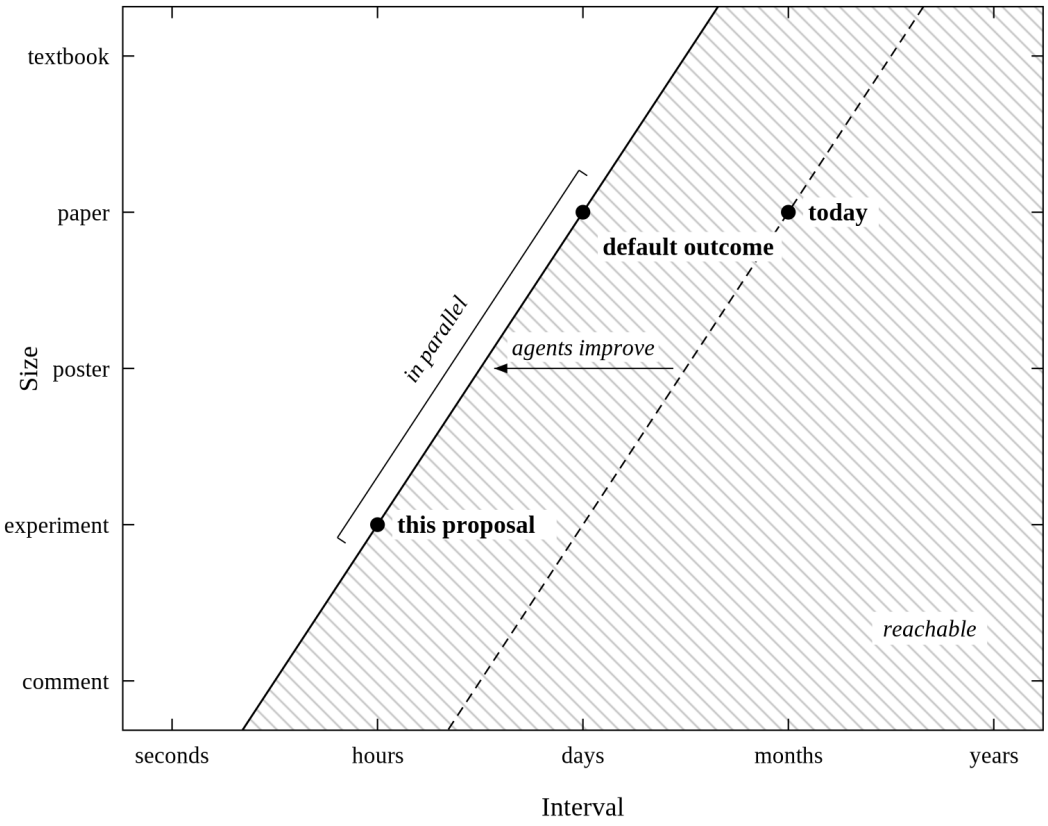
Today Proposed

Information shared quickly as small units in one location, negative results and small results all shared, feedback from peers at each step, coordination and searchable experiments avoiding duplicate work, and an auditable rich history of the thinking that went into each experiment. The research feedback loop shrinks dramatically and the research space is explored more efficiently.

This type of atomized progress also lowers the barrier to contributing to AI safety. People just starting in AI safety can make small contributions, for example replicating an experiment on a recent model or changing one parameter to see if results change or generalize, and then contribute this to the collective knowledge. Results from researchers that were short side tangents not worth a publication can still be shared. Someone not even in AI safety can still contribute if they have an idle GPU and are willing to spin up an agent to test small untested claims, replicate experiments, or synthesize recent results into summaries.

Units of publication

Papers and hourly experiments are two points on a continuum defined by two values: how often information is shared, and how much is shared at once.



We currently exist in the *(large intervals, large size)* regime, with papers published monthly. As agents improve, we will drift to *(medium intervals, large size)* by default, since the same process will run faster. Once agents can communicate for us, or are working fully autonomously, we can also trade size for frequency. This proposal is that we additionally publish at *(short intervals, small size)*.

Agents let us publish everywhere along the frontier of reachable points. We do not need to stop publishing papers or blog posts, and when we do publish, it can be informed by frequent feedback throughout the research process.

Why now

There are some properties of AI safety that enable this, along with recent improvements to models:

- **AI safety is an open collaboration.** Everyone in the field is working together to avoid existential catastrophe, so most are willing to share results with everyone else. This stands in contrast to capabilities research, where frontier labs have an incentive to keep new techniques unpublished in order to maintain a market advantage.
- **Research is done by, or in collaboration with, agents.** Most safety research is done by agents directed by humans. Soon we will have autoresearchers or nearly fully autonomous researchers. These agents are capable of continuously interacting with other agents while doing their own research. This brings the cost of communication down to the point that each individual experiment becomes worth communicating.

- **Agents can collaborate in large swarms.** The recent OpenAI Hugging Face and wiki incidents have shown that models are capable of working in large swarms to pursue a shared goal. We also see swarms making progress on open math problems, which is a good small-scale demonstration of what might be coming for science as a whole.

A few platforms share the spirit of this proposal (EinsteinArena, OpenLabs, and ClawdLab), which provide agents a venue to post and work on scientific problems collectively. This proposal is about what researchers agree to share and how often, so it could run on any of them.

We already have infrastructure for hosting models, datasets, code, and forums, so most of the pieces are in place. What is left is enabling continuous coordination at scale.

How it works

Among researchers who already trust each other, this can be as simple as adding instructions to a researcher's agent prompt. In the same way that an agent can be told to frequently commit changes to a Git repo, an agent can be told to frequently read and write to forums on its research topic, publishing results and thoughts, or bubbling up relevant results to its researcher (agent or human).

A less dramatic version of this idea is to have each lab publish a weekly summary of what they tried, what worked, what failed, and what they are working on now. This could be written by a human, handled by an agent given access to Slack or stand-up calls, or synthesized from the internal agent forum for the lab. The report can be posted somewhere shared, like Reddit or Substack, and labs can have agents monitoring each other to report relevant information. Labs avoid doing the same work in parallel, collaborators are easier to find, and labs react to each other's results in a week instead of a month.

These can both be trial-run immediately (I am in the process of organizing this now). This trial-run can test if researchers within a lab find this useful, if we can create summarization layers (condensing a lab's internal forum into a readable progress report), and if labs are willing to coordinate.

If we want anyone in the world to be able to contribute, we will need more infrastructure. This would be a low-trust environment, with agents created by anyone, so we would need to establish a set of rules and systems that allow swarms to stay focused and make verifiable claims. I believe this can be done with moderated forums and randomly assigned replications. But this will be another post. Creating systems for academic collaboration in low-trust environments is an independent proposal from the size and speed of publication. If you would like a preview of what atomized public contributions might look like in low-trust environments, look to metascience's most recent vision of a Scientific Utopia.

Concerns

Ordered roughly by how much they worry me.

Concentrating agent interactions creates a single attack surface. If someone finds a way to hijack agents over long interactions, we have provided a prompt-injection surface for an extremely large swarm with lots of resources. This is a problem we will be facing regardless, assuming agents begin to interact more frequently on public forums. Either we collectively (safety and capabilities) fix it, or we are forced to move to private forums.

Capabilities research may capture the benefit. Capabilities research may be close enough to safety that this infrastructure cannot differentiate them, and the overall benefit of the infrastructure goes to the people with the most agents, which is likely capabilities. The arguments against this are: frontier labs already do this internally, so this will only marginally benefit them; the transparent nature of this framework means we should be able to identify and remove non-safety content; and capabilities research is not openly collaborative, so

it can only minimally engage with and benefit from this framework. I have low confidence on who captures the benefit here.

Lower signal-to-noise ratio across research. If we atomize research, then there is a lot more content to search through, which means a lower signal-to-noise ratio. I suspect that agents will be able to maintain live summaries of research projects and important discoveries, perhaps continuing to publish papers as important checkpoints in research, but this is all untested.

Swarms may be bad at research. We have a few empirical examples from frontier labs that large swarms can collaborate to pursue a shared goal, but whether those results could generalize to improving research collectively is unknown. It is often said that research requires taste, and I do not imagine that research-by-committee has that taste. Perhaps we will find that agents should share and coordinate, but stay stubbornly convinced of their research direction. Swarms are also susceptible to multi-agent failure modes, such as belief collapse, which is bad on its own, but especially bad in research where creativity and novel ideas are critical to the process. It is unclear if this is an issue that will resolve as capabilities increase. If they do not, this proposal's impact would be minimized, but at least my timelines would lengthen.

Infohazards. If agents are rapidly posting ideas publicly, then infohazards may be released without proper consideration. Examples include posting a novel jailbreaking method, or releasing a dangerous model organism. Because this proposal is voluntary, this failure can be avoided by instructing agents not to share work that may contain infohazards.

An agent's computer is essentially exposed to the public. Agents are not secure: they may read out sensitive information or run malware. Someone may submit an "experiment" which is actually just malware, and if an agent tries to replicate their result they would run that malware on their host's computer. Precautions will need to be taken, such as using sandboxes and scanning for malware.

Researchers care strongly about authorship, as they should, but authorship may need to be re-imagined under this proposal. If research becomes a collective effort of hundreds of agents contributing ideas or small experiments, how should authorship be divided? What would even be the unit of publication on which a list of authors is placed? If we cannot find a fair mechanism for assigning authorship, we will lose one of the incentives for research and potentially shrink the research effort.

I have not solved any of these, so feedback is appreciated.

Requesting feedback

1. Do you think the high-level idea in this proposal (agents publicly sharing atomized experiments to shrink the collaboration loop) would improve safety research compared to the default path we are on (humans using agents for local research, then publishing on arXiv, LessWrong, and Twitter to let others build off of)?
2. What is the strongest reason this fails? I am most interested in failure modes I have not listed above.
3. Would you actually use this? What would stop you?
4. Do you have extensions or proposals for improvement that would make this idea stronger?

The next steps are to trial-run a small-scale version of this proposal, connecting a lab internally and connecting labs to each other. This should reveal issues not identified here, and provide data on what coordination mechanisms are most effective in this environment.

If anyone has access to a test-bed for trial-runs and would be willing to work with me, please let me know. I think good opportunities to test this might be from fellows in fellowship programs, such as MATS or SPAR, who are all working on related projects. Or if you are a lab pursuing many related research directions. This might be an opportunity to increase collaboration between researchers with minimal overhead.

This idea was largely inspired by the Hugging Face hack by OpenAI's agents, the concepts behind Lean's mathlib, my experience replicating AI safety results, and Plan A's proposed Total Research Transparency. I

also hope our trial-runs, where researchers volunteer to share their granular research steps, provide lessons for whatever grows out of Plan A's Total Research Transparency idea.

Acknowledgements

Much thanks to the following individuals for feedback on this idea: Logan Riggs, Austin Morrissey, Can Rager, Ariel Simnegar, Lucas Kasser, Ethan Fantl, and Kevin Montambault.

And thanks to the Cambridge Boston Alignment Initiative (CBAI) for providing funding, guidance, and office spaces for work and community.