The AI Safety Research Fund

Executive Summary

The AI Safety Research Fund is a fiscally-sponsored nonprofit organization dedicated exclusively to funding AI safety research and initiatives. We aim to address critical gaps in the current funding ecosystem by providing a streamlined, transparent, and responsive grant-making process focused solely on ensuring AI has a beneficial impact on humanity.

Funding Gaps

"**The entire workforce preventing AI catastrophe is smaller than two UK high schools**."
-Dewi Erwan, CEO of BlueDot Impact

Funding for AI capabilities is projected to total $360 billion or higher by the end of 2025, yet investments in safety still receive only around $110-130 million annually, creating a funding ratio of 2,800-to-1. Despite a smaller pool of funding, many experts in the field identify severe funding gaps. In a survey of 53 specialists in AI reliability and security research areas, the Institute for AI Policy and Strategy identified 15 sub-areas within AI Safety that were both important (likelihood to significantly reduce AI harms if solved) and tractable (ability for an additional $10M USD to make significant progress on in the next 2 years).

Critical funding gaps identified include Scaling Patterns; CBRN (Chemical, Biological, Radiological, and Nuclear) evaluations; Evaluating Deception, Scheming, Situational Awareness, and Persuasion; and Multi-agent safety. There also remain many Strategic Long-Term Opportunities for projects in AI Safety that were projected to take more than 2 years, but still identified as valuable. This included Supply Chain Integrity and Secure Development, Mechanistic understanding, and Eliciting Latent Knowledge.

Key Differentiators

- **Independence**: No ties to AI companies or political affiliations that could create conflicts of interest. **The majority of current funders for AI Safety have deep connections to major AI labs**. Though they may think they have reasons for this, the outside view is that they act as a revolving door for major AI labs, with their integrity compromised and trust broken with the public.
- **Exclusive AI Safety Focus**: We offer a direct funding channel exclusively for AI safety, eliminating the need to sift through unrelated causes.
- **Curated Funding Opportunities:** We streamline the giving process by identifying and vetting high-impact AI safety projects, removing your research burden.
- **Predictable Decisions & Deadlines:** We commit to clear communication and firm deadlines, ensuring applicants always know when to expect a decision. Our timelines are built on realistic, pessimistic forecasts to avoid delays and allow for better planning.
- **Nurturing Novel & Early-Stage Initiatives:** We actively seek and support a diverse array of projects, including those considered unconventional or nascent, ensuring promising ideas don't go unfunded due to a lack of established history.
- **Seed Funding:** We provide crucial early-stage capital, empowering new AI safety organizations to establish themselves and prove their impact, preventing promising initiatives from dissolving prematurely.
- **Broad-Based Fundraising Strategy:** We proactively seek funding beyond traditional "EA circles" and major donors, engaging a diverse range of individuals and communities to broaden the support base for AI safety.

The Problem We're Solving

Current Funding Landscape Challenges

1. **Limited Options**: Most funding is bundled with other EA/longtermist causes, deterring donors primarily concerned with AI safety
2. **High Friction or limited donor base**: Existing funds often only have a few key donors backing them. The others often require donors to conduct their own research for impact assessment.
3. **Poor Applicant Experience**:
    - Extended wait times (some have averaged 5+ months, despite promising 4 weeks)
    - Psychological harm from perpetual uncertainty
    - Lack of communication and transparency
4. **Underserved Areas**: Established funders often neglect "weird" or early-stage projects
5. **Capacity Constraints**: Major funds operate with zero full-time staff dedicated to AI safety

The Cost of Inaction

The counterfactual to this project is: **nothing happens.** This means less funding will be flowing into AI Safety, fewer alignment researchers will receive a salary, fewer new safety orgs will be created, and fewer prospective researchers will be trained and onboarded.

Time is not on our side. The safety community needs more scalable and responsive funding infrastructure. We're here to build it.

Organizational Structure

Leadership Team

- **Project Lead** (Jaeson Booker): Full-time role overseeing all operations
    - Background: Software engineering, AI safety research, participant of John Wentworth's online training program for MATS, and AI Safety Camp participant
    - Currently dedicating full effort to establishing the fund

Advisory Structure

- **Board of Advisors**: The Advisors help in recommending projects to fund, approving or rejecting grants, and providing clarity on the current state of AI risk and safety.
    - Role: Grant approval/rejection, strategic and research guidance
- **Operational**: Experts in nonprofit management and grantmaking. 1 current member, working part-time but ready to transition to a full-time role.

    - Role: Identifying failure modes, best practices, guidance, compliance.
- **Outreach**: Experts in fundraising, marketing, and outreach.
    - Role: Providing guidance on fundraising initiatives, donor relations, and networking.

Funding Approach

What We Fund

- Individual researchers at all career stages
- New organizations in formation (including pre-incorporation)
- Technical alignment research
- Field-building initiatives
- "Weird" or unconventional approaches often overlooked by traditional funders

Grant Structure

- Small grants: Up to $20,000 (streamlined application)
- General grants: $20,000+ (comprehensive review)

The AI Safety Research Fund will not operate by the normal “impact assessments” or “cost effectiveness” others have used. The “math” behind these assessments is often meaningless and used more as a post-hoc justification for broader, underlying intuitions.

The fund will instead have a kind of “portfolio” of projects and researchers. We will keep up with our projects and researchers to see how they are progressing, and get them any resources we reasonably can (collaboration, references, further funding, etc.) We see our grants as an investment in future outcomes, and we will work to ensure that investment pans out as best as it can. We will also have a diversified portfolio, with the same mentality of many angel investors: high risk, high reward. We will execute intelligent search for potential “unicorns” in alignment research that have gone unnoticed by others. Our primary goal is to seek out and find such research, or even put out a request for the sort of projects we are interested in funding.

The strength of our impact will not be based on the strength of individual projects, but on the overall strength of the portfolio. Research in AI Safety does not and cannot work in isolation, so it is the strength of the whole that will ultimately matter. And the strength of our portfolio will be assessed by its returns: the amount of research that gets done, the number of other researchers who build on it, the new insights gained or understanding that get unlocked, anything that allows our strategic outlook to the future to be in higher resolution and sharper focus.

What we will immediately fund

Below is a list of projects and researchers we are ready to immediately fund with your support.

Aether

About: Aether is an independent LLM agent safety research group. They have $200,000 in funding to cover salaries and expenses through Dec 1, 2025, but do not currently have funding after the next six months.
Team

- Rohan Subramani studied CS and Math at Columbia, where he helped run an Effective Altruism group and an AI alignment group. He has done AI safety research in LASR, CHAI, MATS, and independent groups. He is starting a PhD at the University of Toronto this fall with Prof. Zhijing Jin and continuing Aether work within the PhD.
- Rauno Arike studied CS and Physics at TU Delft, where he co-founded an AI alignment university group. He has done MATS 6 with Marius Hobbhahn, contracted with UK AISI, and worked as a software engineer.
- Shubhorup Biswas is a CS grad and a (former) software engineer, with experience across product and infra in startups and big tech. He did MATS 7.0 with Buck Shlegeris working on AI Control for sandbagging and other low stakes failures.

They are advised by Seth Herd (Astera Institute), Marius Hobbhahn (Apollo Research), Erik Jenner (Google DeepMind), and Francis Rhys Ward (LawZero).

Links: Announcement, Aether July Update

Samuel Gélineau

About: Samuel Gélineau is both a software architect, a Programming Languages Theory (PLT) researcher, and a dad. His most recent publications are about Klister and Möbius, two programming languages with very different ways of combining macro-expansion and type-checking. He is most well-known in the Haskell community for the clarity of his presentations, both at conferences, at local meetups, at McGill's PLT lab (CompLogic), in his online course, and on YouTube. He is working on technical alignment in his spare time, and is looking for funding so that he can focus on this important problem full-time. His approach is to apply PLT to neural networks, to verify that the AI will follow its safety property at runtime in the same way that a type-checker verifies that a program won't fail with a type error at runtime.
Links: AI Safety via Static Analysis, Can we Prove Facts About Machine Learning Models via Code synthesis?

Sandy Fraser

About: Sandy Fraser is a researcher and software engineer with over 20 years of experience solving hard problems in high-tech domains. He has been training AI models since 2016 — well before the current wave — including classifiers for advanced sensors and natural language processing pipelines for risk management. He has developed novel algorithms in fluid dynamics, achieving 10x performance improvements over existing methods and unlocking new manufacturing workflows. His work has won industry recognition, including the 2015 Victorian iAwards for groundbreaking geospatial systems.

After leading teams at Thoughtworks and VPAC Innovations, Sandy is now pursuing technical alignment research for AI safety. His current work explores interpretability and control of advanced AI, including novel selective regularization techniques for representation engineering and new methods for detecting jailbreaks in language models.

Sandy brings a rare combination of deep technical expertise, creative problem-solving, and the ability to translate cutting-edge research into practical systems. He's seeking to apply two decades of building robust, high-stakes systems to the critical challenge of ensuring advanced AI systems remain safe and aligned. His research aims to develop practical techniques that bridge the gap between theoretical alignment work and deployable safety measures.

Work: Selective regularization for alignment-focused representation engineering, Detecting out of distribution text with surprisal and entropy, Concept-anchored representation engineering for alignment

Other projects

There are many who have participated in upskilling programs, who are ready to contribute to AI Safety, but lack the funds. Some have received grants, but many others have not. With the starting donations, many of these promising new researchers can immediately receive funding. We already have made connections with alignment researchers who are ready to recommend highly-skilled candidates. We are ready to begin issuing grants and building a track record. Any donation will make a substantial difference, since every new, highly-skilled researcher might make a substantial difference long term.

Other promising avenues we wish to pursue are new forms of interpretability, such as developmental and representative engineering. Any projects that combine the theoretical perspectives of alignment with the empirical prosaic. Moonshot approaches, such as “Brain based AI” and more conceptual/theoretical work in agent foundations. Multi-agent safety. And societal & structural resilience to adversarial AI, such as d/Acc approaches. We will continue our agentic search for the most promising avenues for alignment, especially anything neglected, overlooked, or missed in the noise. We will continue to correspond with many researchers, founders, and other grantmakers in the field, across a wide variety of perspectives, to gain as much insight as possible and improve our theory of change.

Operations

For Applicants

- **Guaranteed Response Times**: You WILL hear back by the stated deadline
- **Clear Communication**: Regular updates if delays occur
- **Simplified Applications**: Multi-stage process to respect everyone's time
- **Transparent Criteria**: Clear funding priorities and decision factors
- **Post-grant support**: You will be part of our network, and we will aim to do all we can to ensure our investment can flourish. This includes connecting you to whoever you need, recommending your work to others, or making your research visible.

For Donors

- **Multiple Giving Options**: From $10/month subscriptions to major gifts
- **Tax Deductibility**: Full 501(c)(3) status through fiscal sponsorship
- **Impact Visibility**: Regular updates on funded projects
- **Simple Process**: Just donate - we handle the research and vetting

Financial Model

Fundraising Strategy

- **Phase 1**: Secure $25,000 in pledged commitments
- **Phase 2**: Launch with initial grant round
- **Phase 3**: Scale to $1M+ annually through diverse donor base

Fiscal Sponsorship

Donations are tax-deductible through our 501(c)(3) fiscal sponsor, Institute for Education, Research, and Scholarships (IFERS).

Federal Tax ID number (EIN): 72-1585854

DUNS Number: 173120903

Timeline and Milestones

May-June 2025: Secure team and advisory commitments **Completed**

July 2025: Get fiscal sponsorship **Completed**

August 2025: Launch website **Completed**

September 2025: Start accepting donations **Completed**

October 2025: First grants distributed

December 2025: Public impact report

Risk Management

Mitigation Strategies

- **Start small**: Initial focus on small grants to build track record
- **Hard deadlines**: Internal cutoffs for all applicant decisions
- **Transparent operations**: Regular reporting to maintain donor trust
- **Scalable systems**: Infrastructure that can grow from $100K to $10M+

The Path Forward

We will seek to mitigate potential failure modes, and foresee future bottlenecks before they manifest, using previous organizations as datapoints.

Operational bottlenecks

The AI Safety Research Fund will strive to overcome operational bottlenecks. One is that it will not set hard constraints on how much of the funding is dedicated toward it, and by not having a separate fund for paying staff. Another is to make the fund as modular as possible, with Grant Managers, once we reach a certain scale, being delegated to their own subfields. They would be accountable for the applications for that subfield, issuing their own recommendations for approval, with a projected portion of the funds delegated toward that subfield.

Spinoff from fiscal sponsor

As we spinoff from our fiscal sponsor, the Board of Advisors will be transitioned to the Board of Trustees. This way, the Trustees will be individuals who have worked with the fund for 12-24 months, so we will have a good understanding of their commitment and their trustability.

Separate fundraising and grant cycles.

We will separate our different endeavors to different times, so that they can each get our undivided attention. This begins with a time dedicated to fundraising. Once the fundraising round concludes, it will be followed by a time dedicated toward strategic goals given the amount of funding. Then there will be a grant round based on that strategy, where the focus will be on getting high-value applicants.

This might include giving rewards to those who discover promising applicants. And finally a decision round, where we will decide who gets funded.

Applications

We want to make the applications as time economic for all parties involved as possible. This starts with a quick grant application. The goal of the first application is to filter-out the easy-to-determine applicants who we would know from little information would not be a priority for funding. This way, we are not wasting their time either. The next application would require more details, and take more time to fill-out, but these would be applicants we would seriously consider funding. After a certain scale for our organization, we will separate this second stage into two options: one for smaller grants, such as those under 10k, which would not need as many details. And one option for larger grants, which would need more scrutiny and more details. It's not a one-size-fits-all, and we'll finetune this process as we go along.

Fundraising initiatives

We plan to try a variety of different fundraising tactics, see which ones bare fruit, and then scale them to larger campaigns. This can range from social media campaigns, grassroots organization, targeting high-income donors, leveraging already existing networks, academic circles, and fundraising events. Another goal is for us to be ready for a potential sudden increase in funding AI Safety projects that we will quickly be able to capitalize on.

Donor Interface

Another idea we will be working on is creating a User Interface that will make the AI Safety Landscape understandable and directly usable for donors. This will involve breaking-down Alignment into various easily-understandable subdomains, with various researchers we're interested in funding in those subdomains. Donors will then have the option to donate directly to that researcher or subfield, giving them the knowledge that their contribution is directly useful, while also making it easy-to-use and understandable for laypeople. A small, demo version can be found here.

Funding Opportunities

While there are several funds currently focused on AI Safety, we believe there are gaps in their ability to raise funds from a broader range of donors and allocate grants to all promising fields of research.

Open Philanthropy & Survival and Flourishing Fund: Both organizations have had extraordinary impact in the AI Safety space, providing millions of dollars to highly impactful researchers and organizations that otherwise would not have received funding. However, both are primarily funded by a very small number of donors. Since AI Safety is a very complicated field, with many conflicting perspectives, this likely has the effect of narrowing the range of research funded and leaving critical avenues unexplored. Either explicitly or implicitly, there is

likely less diversity in the scope of projects funded as a result, increasing the probability that high-impact research avenues are being neglected. These organizations are also unlikely to be able to fully capitalize on the growing interest from smaller donors, since outreach and fundraising are not their main objectives. For instance, a standard software engineer, making 100-300k a year, does not have an easy and obvious way to donate to AI Safety.

Longterm Future Fund & Emerging Challenges Fund (Longview): Though both have the option to donate on their websites, their focus on longtermism likely dissuades many potential donors. Another problem is that these funds are not exclusively directed at the risks from AI, including other causes like biosecurity and nuclear weapons. They do not have an option for people who are solely concerned about short-term risks from AI.

Other funds/ventures: There are a number of possible new funds and ventures in the funding space for AI Safety, but they currently have not manifested in useful ways for most donors. The AI Risk Mitigation Fund (ARM) has not fully gotten off-the-ground, despite being in the works for several years, with no public grant rounds, and seemingly no public outreach campaign to reach the sort of people we are trying to reach. The AI Safety Fund was primarily funded by major AI labs, which hurts trust, and has recently lost their fiscal sponsor (Meridian Prime), making their future ability to issue grants currently uncertain. Manifund requires too much research understanding and work to be able to attract a wide donor base. There have been numerous attempts such as these over the years, but as of writing this, none have materialized in a way that meaningfully attracts the donors we are trying to attract.

The Stakes

Funding gaps and neglected research matter much more if the stakes are high. With stakes as high as AI Risk, any gaps in funding might be critically important to address. Any promising research that gets neglected could make the difference between a world where humanity survives and flourishes, and one where we fall into disempowerment or extinction. We have identified a pathway to address this, one that has already garnered interest among many prominent VCs who are interested in donating to AI Safety. We believe this venture can be achieved without your assistance. The difference you can make is in the speed and ease we are able to set up operations, issue the first grants, and build our track record and reputation. Given the uncertainty around AI timelines, achieving this just a few months faster than we otherwise would could make an enormous difference.

Call to Action

Join us in creating the focused AI safety funding mechanism the field desperately needs. Every contribution, from $10/month to major gifts, directly accelerates safety research. This could be the most impactful contribution you have ever made. We are completely committed to make AI Safety succeed, and will use all our efforts to seek out new and promising research, hear new perspectives, and identify unforeseen risks.

Team

Jaeson Booker
Fund Manager

Jaeson is a software engineer and AI alignment researcher with a deep commitment to reducing existential risk from advanced AI. With a background spanning blockchain systems, decentralized governance, and AI research, he brings a rare combination of technical depth and strategic vision to the field.

Contact: jaesonbooke@gmail.com, LinkedIn

Kabir Kumar
Kabir is the Director and Cofounder of AI-Plans.com. His work has made multiple discoveries, helped a team get accepted at an ICML workshop, and has organized AI Safety events with participants from AMD, Meta, and AWS. His expertise in AI strategy and planning helps guide our mission and grant allocation decisions.

Dr. Jobst Heitzig

Jobst is a Senior Researcher for FutureLab on Game Theory and Networks of Interacting Agents at PIK, and Project Lead for SaisfIA. Jobst is a mathematician working on collective decision making, game theory, formal ethics, and international coalition formation.

Seth Herd

Seth spent 23 years in cognitive psychology and neuroscience research, increasingly concerned with its applications to human-like AI. He now researches alignment and AI risk full-time, focusing on the potential for expanding LLMs into loosely brainlike cognitive architectures capable of independent thought and innovation.

Cole Wyeth

Cole is a mathematician and a computer scientist. Just as electrical engineering applies the understanding of electricity and magnetism to build physical artifacts, he uses computer science to apply the abstract formal reasoning of mathematics to build software objects. He holds an M.S. in mathematics from the University of Minnesota, and is a PhD student in computer science at the University of Waterloo. His main interests are trying to align an AIXI agent.

Pip Foweraker

Pip Foweraker advises on AI safety, governance, and operations. As former COO of Ashgro, he scaled operations across dozens of sponsored projects worldwide and collaborated with funders, researchers, and policymakers. His current work explores resilient research infrastructure and national-level AI safety strategy.

Esben Kran

Esban is a research entrepreneur, focused on technical approaches to make AI safer, increasing awareness about the problems, and inspiring the next generation of researchers and hackers. He is co-founder of Seldon Labs and the non-profit research institute Apart Research, and sits on the board of the European Network for AI Safety. He is also an advisor for AI Safety Asia and Juniper Ventures. Esban will join the AI Safety Research Fund as an Advisor once the fund reaches $1 million in donations.

Next Steps:

1. Review this prospectus
2. Donate to our organization via Paypal, Zelle, or ACH.

   OR

   Schedule a follow-up conversation to discuss specifics

Donors who pledge $5000 or greater will have the option to be featured on our website.

We recognize this initiative's critical, time-sensitive need. We invite you to join us. Our existence directly translates to increased funding, better compensated researchers, the creation of new safety organizations, and the training of future talent, all vital for a positive AI future. We are the missing link in the AI safety funding ecosystem.
Let's ensure the future of AI is safe, sane, and human-aligned.
Help us fund it.

Donate Now at https://www.aisafetyfund.org/