

Sponsorships

Team Size organising the program

11 people, all volunteers

Existing Partnerships,

Prev: HUD Evals, ai evals startup

AI Safety Global, non profit

University of Victoria AI Society - student group

Interested and said interested in paying \$20 for a Job Booth - Harmony Intelligence - AI Security startup

Competitors (if any),

No one else is really going research programs specifically targeting the hard part of alignment.

For multi week research programs in ai alignment/safety:

MATS, MARS, SPAR, potentially some others

Why it is unique

focused on the hard part of alignment

crazy talent pool - 50% or so have phds, multiple senior swes and other researchers/engineers from Nvidia, AMD, AWS, Meta, Microsoft, Siemens, etc

Not mentor limited - made very actionable and useful Research Guides, thanks to doing lots of calls/interviews with senior researchers and then doing own study/research into the topics.

So we can support more people.

Strong distribution - Prev 2 events had 150+ signups each, this one has had 298 applications and we can very easily get a lot more - there's a lot of marketing things that we didn't get time for, due to research prep this time - next time, we'll already have the research prep from this, so we'll be able to market more and have even more of a userbase/audience to send to.

- also, sent out official announcement very late, due to spending a lot of time on research prep, but still got this number. had to buy a Notion plan, due to how many people were applying

The quality of the research will be extremely high due to the very tight focus and the amount of research prep we've done ourselves and the quality of people applying. we keep this high quality of people applying each time, due to having a strong research background and actively reading research and learning ourselves, so knowing what is actually the most important and neglected research areas - this then attracts high quality engineers and researchers who want to work in this area but don't have a place to do so.

WE ARE NOT MENTOR LIMITED

- due to the research guides and doing all this research prep, we are not limited by:
 - the quantity of the mentors
 - the pre-existing biases/assumptions of mentors, limiting/stagnating the kinds of research that can be explored

Who are the stakeholders/people interested

engineers/researchers who want to work on these problems

people who care about ai safety/alignment - mostly a bunch of smart rich people, government AI security/safety institutes, alignment researchers

general ai engineers/researchers - it's a very interesting problem in general

neuroscientists who are also interested in the human morality/brain problem

maybe companies who want to hire talented people?

Size of impact being created

prev 1 week hackathon led to multiple teams working together on the problem of alignment evals, and then continuing to work together over a month after the hackathon ended.

currently there are less than 100 people in the world working on the hard part of alignment - 200 if you stretch the definition a bit.

if at the minimum, we have the same number of teams as last time, 2, that would mean 10 new people working on this. that would be a 110% difference in a very important problem.

Given that this is a 5 week program, rather than a 1 week event, where people have more time to get to know each other and enjoy working together, we might have over double that amount - which would mean 20% change in the number of people working on this.

We will also be making Research Guides on how to actually do the research on the hard part of alignment and then releasing those, to an audience of over 700 people, most of whom are ai researchers or interested in ai research.

In particular, we are essentially creating 2 new important research directions - Neuromorality based Alignment and Applied Agent Foundations, both of which are very attractive to researchers/engineers, can be worked on, measured, made progress in and will be very good for alignment.

We're also massively increasing the foundation that Theoretical Agent Foundations has and introducing the topic to IMO winners, Kaggle Grandmasters and multiple Math and Physics PhDs.

Long term plans

Continue hosting program like this, become more and more knowledgeable about the research, make the research guides better and better, increase the quality of the alignment research being done.

On the Poster Day - so that lots of people can present and share their work at the same time.
If we get over \$2000 in fiscal donations, then happy to sponsor free tickets for some people
(each person on GatherTown costs us \$3)

In order of priority:

If we get more, then paying myself (Kabir) a salary of \$500 or \$1000, depending on how much, paying other people who also helped a lot with this:

e.g. Lenz, who made the Notion page, and made it professional, from my extremely messy notes,

Luna who made the design for <https://aiplans.org/events/moonshot-alignment-program>, which looks really professional,

Karan who coded the frontend for that, based on Luna's design

Purva, who's been helping a lot on organizing and making the second Notion page for Participants, so that people can actually take part

Leo and Kabee, who helped share/market the event

- leo helped make the software I used to go through applicants much more quickly
- Leo and Kabee both helped make the Research Guides

Jitesh, who helped make the Research Guides

Laurie, who helped make the Research Guides

Felix, who helped make the Research Guides

David, who helped a bit in organizing the Research Assistants table

<https://aiplans.org/news> - this will be where we make announcements of the results and outcomes.

we also have a sponsorship deal with a deep learning youtuber with 28k subscribers, for \$300:

<https://www.youtube.com/@deeplearningexplained>

he's got a neuroscience phd and is going to be making a video about a neuromorality paper and talking about the Demo Day: <https://lu.ma/8ehdokx7>

Marketing

Messages

Neuroscience Researchers

Hi __, I saw you paper __ . I'm reaching out because I want to know what you think are the current major bottlenecks on figuring out exactly how values/preferences are encoded in the human brain. E.

Recruit Organizers

Hi, I'm hosting the Moonshot Alignment Program on July 26th, which will be a 5 week research program to get there to be more teams trying to solve the hard part of alignment, so we make progress there as well, rather than just sub problems

I'm looking for folk to help organize it

I'm confident we can get a lot of people - prev 2 alignment evals hackathons had 150+ people each sign up

Multiple teams take part, in the last hackathon in April/May, several teams made a full on paper within a week

James Hindmarch, a guest judge, said he was very impressed with the quality of work given the time and resource constraints

Multiple teams from there are still working together now, with one submitting their work to the canadian national security council and another having finished their paper, submitted to arxiv and now starting work on a new eval, on alignment faking

We've also secured a deal with a youtuber with 25k subscribers, Deep Learning With Yacine, who makes technical deep learning videos, who's agreed to make a full video about the research program

Just today someone talked to the dean of their university about the program and he said she was very interested and might email the professors herself - and more students have said they're going to be talking to the deans of their university as well

I've interviewed lots of agent foundations researchers to make sure that what we do in this program will be valuable, including Vanessa Kosoy, Norman, Cole, Plex, etc.
I've also talked with Abram Demski about what he considers to be important about the Tiling Problem

Next planned steps are interviewing more neuroscience researchers (Yacine has a phd in neuroscience, asked him about the neuroscience track in our call as well) and preference

optimization researchers (Ana, a member of the AI Plans team, did her thesis in preference optimization)

Preference Optimization Experts

Subject: non shallow Preference Optimization

Dear __,

I'm __, I'm helping organize a research program on making and improving preference optimization architectures. I am reaching out to you because of your work in preference optimization.

We're currently organizing a [5 week research program](#) focused on building and improving AI architectures that encode values in a way that generalizes robustly, scales gracefully, and reflects meaningful internal representations, rather than surface-level behavioural patterns or shallow refusal mechanisms.

As part of this effort, we're speaking with researchers like yourself to better understand the current landscape and learn what are the most useful problems for researchers to work on.

This is both to make a guide for participants on how to do valuable research during the program and to know how to best structure the program so that the most important problems get targeted.

Depending on your availability, we'd love to either:

- [Have a call](#) to learn the specific bottlenecks of your research, what you think good/bad research looks - this would be with the head researcher, Kabir
 - We would use this to make a detailed guide on how to have good ideas for the research and mistakes to avoid.
- Invite you to give a talk during the program, about your research
- Let you select a team of mentees for a research project of your choice - with us providing compute credits, custom evals, etc

Looking forward to hearing from you,

AI Plans

Thing to change in here - make it shorter

make them feel more that we're interested in them and want to be assets to them, rather than wanting to use them as an asset - saying like 'i want to help your research go faster - whats the current major bottlenecks you have'

for Kabir

I'm Kabir Kumar, I'm running an alignment research team. I am reaching out to you because of your work in preference optimization.

We're currently organizing a [5 week research program](#) focused on building and improving AI architectures that encode values in a way that generalizes robustly, scales gracefully, and reflects meaningful internal representations, rather than surface-level behavioural patterns or shallow refusal mechanisms.

As part of this effort, we're speaking with researchers like yourself to better understand the current landscape and learn what are the most useful problems for researchers to work on.

This is both to make a guide for participants on how to do valuable research during the program and to know how to best structure the program so that the most important problems get targeted.

Depending on your availability, we'd love to either:

- [Have a call](#) to learn the specific bottlenecks of your research, what you think good/bad research looks - this would be with the head researcher, Kabir
 - We would use this to make a detailed guide on how to have good ideas for the research and mistakes to avoid.
- Invite you to give a talk during the program, about your research
- Let you select a team of mentees for a research project of your choice - with us providing compute credits, custom evals, etc

Looking forward to hearing from you,

AI Plans

Dean/Department Chair

- this is a legit research program
- we're doing useful research
- if their students were to do this research, they would have high chance of publishing
- a previous team from a hackthon is presenting at ICML
- multiple teams from hackathons have continued working together on research
- many said they found new and really interesting things to do because of it
- give a vibe that this is prestigious and would elevate the prestige of their department/university without actually saying so - for deans/department chairs - not professors
- for professors - mainly that students really like it, it's useful research

Email Subject: Research Program on AI Opportunity

Dear Dean [Name],

We hope this message finds you well. We are writing to introduce a research opportunity that may be of significant interest to your students and professors.

AI Plans is pleased to announce our upcoming five-week intensive research program called [Moonshot Alignment Program](#) focused on developing advanced architectures for preference optimization and value alignment in artificial intelligence systems.

Teams in previous AI Plans events have made novel discoveries, such as LLMs being maximizers in long horizon multi-tasking scenarios, with some presenting at the International Conference on Machine Learning (ICML). The program will culminate in a virtual poster day and a job fair. An expert panel will vote on standout projects, while research organizations/labs host booths and share open roles.

Participant feedback consistently indicates that this program provides one of the most intellectually stimulating and innovative research experiences available in the field. The program combines rigorous theoretical foundations with practical applications, offering students the opportunity to work on cutting-edge problems in AI safety and alignment under expert supervision.

We believe that students and professors from your institution would make valuable contributions to this research community, and we are confident that their participation would reflect positively on the excellent research environment and academic standards that your department maintains.

There program will have five tracks to choose from:

1. Agent Foundations Theory - solving complex math problems in computational learning theory and theoretical computer science.
2. Agent Foundations Applied - making new architectures based on existing agent foundations research.

3. Moral Cognition Based Architectures - taking what is known about moral cognition and preferences in neuroscience to experiment with new training methods, to try to create new, robust, and scalable value learning methods.
4. Improved Preference Optimization Methods - improving upon works such as PPO, DPO, KPO, AI Safety via Debate, etc., to make new, non-shallow preference optimization methods.
5. Original/Others - this is an open track for methods which don't fit into the above categories, but still show promise.

The target audience for this event is those with a background in: pure mathematics, theoretical computer science, neuroscience, preference optimization, or ML architectures. Our previous events had more than 150 signups, with participants including an AMD engineer, multiple quants, PhDs, and Wall Street ML engineers.

You can find more details about the program and the registration in this link [Moonshot Alignment Program](#).

Anyone is welcome to apply, but we have limited spots due to the amount of mentors available.

Thank you for your time and consideration. We look forward to the possibility of more collaborations especially in supporting the academic development of your students.

Sincerely,

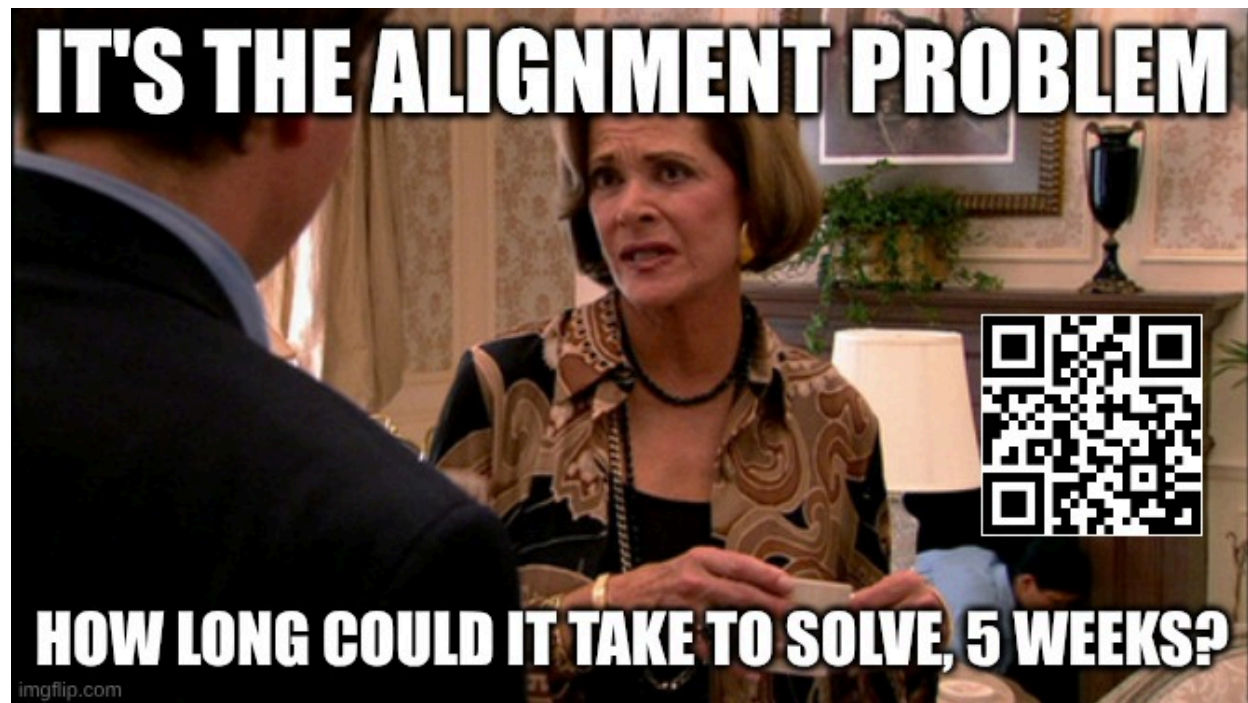
AI Plans

<https://ai-plans.com>

[Contact Email]

[Contact Phone Number]

for Groups



I'm hosting a 5 week research program on directly tackling the hard part of alignment!

<https://courageous-lift-30c.notion.site/Moonshot-Alignment-Program-20fa2fee3c6780a2b99cc5d8ca07c5b0>

First 300 applicants are guaranteed personalized feedback! (94 applied so far)

Deadline to apply: 25th July

Message to University Heads

University heads

Dear (name),

AI Plans are hosting a 5-week research print from August 2nd focused on creating new, robust, and scalable methods to encode values in AI models. The research tracks include: agent foundations (theoretical & applied), brain-based AI safety, improving preference optimization methods, and original/other methods.

Teams in previous AI Plans events have made novel discoveries, such as LLMs being maximizers in long horizon multi-tasking scenarios, and that existing safety evaluations (before January 2025) can easily be gamed if the model is behind an API. Teams have also gone on to present at ICML and to the Canadian National Security Council. The program will culminate in a virtual poster day and a job fair. An expert panel will vote on standout projects, while research organizations/labs host booths and share open roles.

The basic program structure is as follows:

Week 1 (beginning August 2nd): form teams, define specific approaches, design falsifiable experiments

Weeks 2-3: build implementations, iterative testing at increasing scales

Week 4: critique method, attempt to red-team evals, document failure modes and anomalies

Week 5: write research summary, prepare poster and get final feedback

The target audience for this event is those with a background in: pure mathematics, theoretical computer science, neuroscience, preference optimization, or ML architectures. Previous events both had 150+ signups, with participants including an AMD engineer, multiple quants, PhDs, and Wall Street ML engineers.

Here's a Notion

page(<https://courageous-lift-30c.notion.site/Moonshot-Alignment-Program-20fa2fee3c6780a2b99cc5d8ca07c5b0>) with more details.

If you'd like to have a call, here the Calendly of Kabir Kumar, CEO of AI Plans (<https://calendly.com/kabir03999/talk-with-kabir>).

I have included a template email that you can send to students below.

Kind regards,

Ben

Volunteer at AI Plans (<https://ai-plans.com/>)

--

Beginning on August 2nd, the Moonshot Alignment Program is a 5 week fellowship on creating new, robust and scalable methods to encode values in AI models.

Apply

here(<https://courageous-lift-30c.notion.site/21ba2fee3c6780a4864bd6aa992ff35a?pvs=105>) before the 20th of July.

There will be 5 tracks to choose from:

Agent Foundations Theory—solving complex math problems in computational learning theory and theoretical computer science

Agent Foundations Applied—making new architectures based on existing agent foundations research

Moral Cognition Based Architectures—taking what is known about moral cognition and preferences in neuroscience to experiment with new training methods, to try to create new, robust, and scalable value learning methods

Improved Preference Optimization Methods—improving upon works such as PPO, DPO, KPO, AI Safety via Debate, etc., to make new, non-shallow preference optimization methods

Original/Other—this is an open track for methods which don't fit into the above categories, but still show promise

Teams in previous events by AI Plans have presented at ICML, made new, novel research methods, and important discoveries such as LLMs being maximizers when balancing multiple tasks. The Moonshot Alignment Program will cumulate in a virtual poster day—where teams will have a chance to present their work to up to hundreds of researchers—and a job fair, where companies will be sharing roles they're looking for.

Anyone is welcome to apply, but we have limited spots due to the amount of mentors available. If you have a background in pure mathematics, theoretical computer science, neuroscience, preference optimization, or ML architectures, we encourage you to apply. Research participants are expected to deliver 10 hours per week.

See

here(<https://courageous-lift-30c.notion.site/Moonshot-Alignment-Program-20fa2fee3c6780a2b99cc5d8ca07c5b0?pvs=74>) to learn more.

Subject: Research Program Opportunity - Moonshot Alignment Program

Dear [University Name],

We hope this message finds you well. We are writing to introduce a research opportunity that may be of significant interest to your students and faculty members.

AI Plans is pleased to announce our upcoming five-week intensive research program, the [Moonshot Alignment Program](#), focused on developing advanced architectures for preference optimization and value alignment in artificial intelligence systems.

Teams participating in previous AI Plans events have made significant novel discoveries, including findings regarding LLMs as maximizers in long-horizon multi-tasking scenarios. Our alumni have presented their research at prestigious venues such as the International Conference on Machine Learning (ICML) and to the Canadian National Security Council. The program culminates in a virtual poster presentation and job fair, where an expert panel evaluates standout projects while research organizations, companies and laboratories host booths to share available positions.

Participant feedback consistently indicates that this program provides one of the most intellectually stimulating and innovative research experiences available in the field. The program combines rigorous theoretical foundations with practical applications, offering participants the opportunity to work on cutting-edge problems in AI safety and alignment under expert supervision.

The program offers five specialized research tracks:

1. Agent Foundations Theory - addressing complex mathematical problems in computational learning theory and theoretical computer science.
2. Agent Foundations Applied -developing new architectures based on existing agent foundations research.
3. Moral Cognition Based Architectures - applying neuroscience insights about moral cognition and preferences to experiment with novel training methods and create robust, scalable value learning approaches.
4. Improved Preference Optimization Methods - advancing beyond existing methods such as PPO, DPO, KPO, and AI Safety via Debate to develop sophisticated, non-shallow preference optimization techniques.
5. Original/Others - an open track for innovative methods that demonstrate promise but do not fit within the above categories.

This program targets individuals with backgrounds in pure mathematics, theoretical computer science, neuroscience, preference optimization, or machine learning architectures. Our previous events have attracted over 150 participants, including AMD engineers, quantitative analysts, PhD holders, and Wall Street machine learning engineers.

We believe that students and faculty from your institution would make valuable contributions to this research community, and we are confident that their participation would reflect positively on the excellent research environment and academic standards that your institution maintains.

Detailed program information and registration are available at the [Moonshot Alignment Program](#) website. While applications are open to all qualified candidates, spaces are limited due to mentor availability.

We would be grateful if you could share this opportunity with your students and faculty through your department's announcements or appropriate channels.

Thank you for your time and consideration. We look forward to potential collaborations and supporting the academic development of your students and faculty.

Sincerely,

AI Plans

<https://ai-plans.com>

[Contact Email]

[Contact Phone Number]

Company Messages

Hi, I'm offering a chance to recruit from a group of PhDs, Quants, AI Engineers and Researchers.

We're hosting a Virtual Job Fair on September 6th, as part of a research event, where AI engineers and researchers from AMD, Oxford, JP Morgan Chase and other companies are taking part

We're offering Job Booths, where you will be able to recruit and advertise to these researchers, for \$200.

Subject: Moonshot Alignment Program Sponsorship Opportunity

Dear [Company Name],

We hope this message finds you well. We are writing to present an exciting sponsorship opportunity that aligns with your organization's commitment to advancing artificial intelligence research and talent acquisition.

AI Plans is pleased to announce our upcoming five-week intensive research program, the [Moonshot Alignment Program](#), focused on developing advanced architectures for preference optimization and value alignment in artificial intelligence systems. This program represents a unique opportunity for your organization to connect with top-tier AI talent and showcase your commitment to cutting-edge research.

Program Impact and Reach

Teams participating in previous AI Plans events have made significant breakthrough discoveries, including novel findings regarding LLMs as maximizers in long-horizon multi-tasking scenarios. Our alumni have presented their research at prestigious venues such as the International Conference on Machine Learning (ICML).

Our previous events have attracted over 500 highly qualified participants, including engineers from leading technology companies, quantitative analysts, PhD holders, and machine learning specialists from major financial institutions. This program specifically targets individuals with expertise in pure mathematics, theoretical computer science, neuroscience, preference optimization, and ML architectures.

Sponsorship Opportunities

We offer two strategic sponsorship packages designed to maximize your organization's visibility and recruitment potential. You can avail of both packages as well.

Virtual Booth Sponsorship - \$200

- Dedicated virtual booth space during our concluding job fair at Gathertown.
- Direct access to program participants for networking and recruitment.

- Company logo and information displayed prominently in communication materials and during the event.
- Opportunity to share open positions and internship opportunities in the job fair.
- Access to participant profiles and research outputs.

Premium Speaking Engagement - \$2,000

- 15-minute virtual presentation slot during the program at Gathertown.
- Virtual booth inclusion (full benefits listed above).
- Priority placement in program communications.
- Enhanced logo visibility and information across all program materials.
- Opportunity to present technical challenges or research directions to participants after the event.

Strategic Value Proposition

This sponsorship provides your organization with:

- Direct access to emerging talent in AI safety and alignment research.
- Brand visibility among highly skilled researchers and academics.
- Opportunity to influence the next generation of AI researchers.
- Platform to showcase your organization's technical challenges and innovations.
- Cost-effective recruitment channel for specialized AI talent.

The program's focus on preference optimization and value alignment positions which your organization at the forefront of critical AI research areas will find important especially with the increasing demand for responsible AI development.

We would welcome the opportunity to discuss how this sponsorship aligns with your organization's goals and recruitment needs. Please feel free to contact us to arrange a brief call to explore the partnership possibilities. Space is limited for both sponsorship opportunities, and we anticipate strong interest from organizations seeking to connect with this exceptional talent pool.

Thank you for considering this opportunity to support cutting-edge AI research while advancing your organization's talent acquisition and helping in the development of safe and responsible AI.

Sincerely,

AI Plans

<https://ai-plans.com>

[Contact Email]

[Contact Phone Number]

For compute sponsorship

Subject: Moonshot Alignment Program Sponsorship Opportunity

Dear [Company Name],

We hope this message finds you well. We are writing to present an exciting sponsorship opportunity that aligns with your organization's commitment to advancing artificial intelligence research and talent acquisition.

AI Plans is pleased to announce our upcoming five-week intensive research program, the [Moonshot Alignment Program](#), focused on developing advanced architectures for preference optimization and value alignment in artificial intelligence systems. This program represents a unique opportunity for your organization to connect with top-tier AI talent and showcase your commitment to cutting-edge research.

Program Impact and Reach

Teams participating in previous AI Plans events have made significant breakthrough discoveries, including novel findings regarding LLMs as maximizers in long-horizon multi-tasking scenarios. Our alumni have presented their research at prestigious venues such as the International Conference on Machine Learning (ICML).

Our previous events have attracted over 500 highly qualified participants, including engineers from leading technology companies, quantitative analysts, PhD holders, and machine learning specialists from major financial institutions. This program specifically targets individuals with expertise in pure mathematics, theoretical computer science, neuroscience, preference optimization, and ML architectures.

Sponsorship Opportunities

We offer the following strategic sponsorship packages designed to maximize your organization's visibility and recruitment potential. You can avail of any or all of the packages as well.

Virtual Booth Sponsorship - \$200

- Dedicated virtual booth space during our concluding job fair at Gathertown.
- Direct access to program participants for networking and recruitment.
- Company logo and information displayed prominently in communication materials and during the event.
- Opportunity to share open positions and internship opportunities in the job fair.
- Access to participant profiles and research outputs.

Premium Speaking Engagement - \$2,000

- 15-minute virtual presentation slot during the program at Gathertown.
- Virtual booth inclusion (full benefits listed above).
- Priority placement in program communications.

- Enhanced logo visibility and information across all program materials.
- Opportunity to present technical challenges or research directions to participants after the event.

Compute and Discount Sponsorships

- Provide compute or subscriptions discount credits to at least the top 10 to 50 research teams.
- Virtual booth inclusion (full benefits listed above).
- Priority placement in program communications.
- Enhanced logo visibility and information across all program materials.

Strategic Value Proposition

This sponsorship provides your organization with:

- Direct access to emerging talent in AI safety and alignment research.
- Brand visibility among highly skilled researchers and academics.
- Opportunity to influence the next generation of AI researchers.
- Platform to showcase your organization's technical challenges and innovations.
- Cost-effective recruitment channel for specialized AI talent.

The program's focus on preference optimization and value alignment positions which your organization at the forefront of critical AI research areas will find important especially with the increasing demand for responsible AI development.

We would welcome the opportunity to discuss how this sponsorship aligns with your organization's goals and recruitment needs. Please feel free to contact us to arrange a brief call to explore the partnership possibilities. Space is limited for both sponsorship opportunities, and we anticipate strong interest from organizations seeking to connect with this exceptional talent pool.

Thank you for considering this opportunity to support cutting-edge AI research while advancing your organization's talent acquisition and helping in the development of safe and responsible AI.

Sincerely,

AI Plans

<https://ai-plans.com>

[Contact Email]

[Contact Phone Number]

Shortened version

Subject: Moonshot Alignment Program Sponsorship Opportunity

Dear [Company Name],

AI Plans is pleased to announce our upcoming five-week intensive research program, the [Moonshot Alignment Program](#), focused on developing advanced architectures for preference optimization and value alignment in artificial intelligence systems. This program represents a unique opportunity for your organization to connect with top-tier AI talent and showcase your commitment to cutting-edge research.

Program Impact and Reach

Our previous events have attracted over 500 highly qualified participants, including engineers from leading technology companies, quantitative analysts, PhD holders, and machine learning specialists from major financial institutions. This program specifically targets individuals with expertise in pure mathematics, theoretical computer science, neuroscience, preference optimization, and ML architectures.

Sponsorship Opportunities

We offer the following strategic sponsorship packages designed to maximize your organization's visibility and recruitment potential. You can avail of any or all of the packages as well.

Virtual Booth Sponsorship - \$200

- Dedicated virtual booth space during our concluding job fair at Gathertown to share open positions, internship opportunities and have virtual meetings with participants.
- Direct access to program participants for networking, research outputs and recruitment.
- Company logo and information in communication materials and during the event.

Premium Speaking Engagement - \$2,000

- 15-minute virtual presentation during the program at Gathertown and all virtual booth inclusion.
- Priority placement in communications and information across all program materials.
- Opportunity to present research directions to participants after the event.

Compute and Discount Sponsorship

- Provide compute or subscriptions discount credits to at least the top 10 to 50 research teams.
- Virtual booth sponsorship inclusion benefits and opportunity to present research directions to participants after the event.

Strategic Value Proposition

This sponsorship provides your organization with:

- Direct access to emerging talent in AI safety and alignment research.
- Brand visibility among highly skilled researchers and academics.

- Opportunity to influence the next generation of AI researchers.
- Platform to showcase your organization's technical challenges and innovations.
- Cost-effective recruitment channel for specialized AI talent.

The program's focus on preference optimization and value alignment positions which your organization at the forefront of critical AI research areas will find important especially with the increasing demand for responsible AI development.

We would welcome the opportunity to discuss how this sponsorship aligns with your organization's goals and recruitment needs. Please feel free to contact us to arrange a brief call to explore the partnership possibilities. Feel free to book a 1on1 call at this Calendly link. Space is limited for both sponsorship opportunities, and we anticipate strong interest from organizations seeking to connect with this exceptional talent pool.

Thank you for considering this opportunity to support cutting-edge AI research while advancing your organization's talent acquisition and helping in the development of safe and responsible AI.

Sincerely,

AI Plans

<https://ai-plans.com>

Groups to Reach

Ai Control

Ai security

ARENA based program

Neuromatch

AI Safety Group

Event Prep

Notion Page

Flow:

Multiple Choice Checkboxes, putting their background/experience:

then show them the paths/tracks that would be most relevant for them, in order

e.g. Neuroscience PhD, Preference Optimization, Agent Foundations Applied, Agent Foundations Theory

Schedule

When they can expect to hear back about stage 2 or feedback/questions or rejection from Stage 1

If rejected Stage 1, encourage to take part in Critique-a-Thon and Alignment Evals hackathon, which are open to everyone

Apologise for rejection, say due to this being a 5 week program, we don't have the resources to accomodate everyone, as we'd like to

Maybe suggest some learning resources/exercises for the track they chose but didn't get into

e.g. if they applied for agent foundations theory - something they can do to strengthen future applications is solve a proof/problem in computational learning theory

if neuroscience track, take brain data, maybe use new methods that havent been applied yet, to learn more about how preferences are encoded in the brain, or find ways to improve data/signal to noise ratio from fmeg and other brain scans

if preference optimization track, writing about specific reasons current methods such as constitutional ai dont work robustly or at scale and what might be needed

in general, making a very good alignment eval or red teaming alignment evals is good -

- for this, we should take the explainer doc for evals we have and turn that into self contained, by itself, explainers for how to make a good alignment eval and how to red team an alignment eval

Interviews

Agent Foundations Researchers

Agent Foundation is the 'everything else'

- come in with fewer assumptions

"we'd like these properties"

Paul Rapoport

Bold = You need to know to start working on this

General helpful prereqs/competencies:

- Measure theory/real analysis
- Propositional logic
- "Intro to proofs"/being able to construct a solid mathematical argument
- Being able to construct a workable informal mathematical argument-sketch
- Bayes nets/causal DAGs
- Being able to quickly learn new math
- Operationalization skills - expressing philosophical objects robustly in math
- For some directions it's useful
- These are very basic core competencies

(The below is probably very very incomplete)

Infrabayes

- **Measure theory/real analysis**
- Convex optimization (probably)
- **Brushing up specifically on the new foundations for IB** (will need to ask Vanessa; AIUI she hasn't published it publicly yet)
- Learning theory/(PO)MDPs

UDT/"Tiling Agents"

- Lots of basic logic (**propositional logic**)
- UDT-specific background (timelessness/updatelessness, logical inductors, logical induction markets)
- Computability/provability theory
- Jeffrey-Bolker prob-utility framework? [Probability is Real, and Value is Complex — LessWrong](#) [Subjective Naturalism in Decision Theory: Savage vs. Jeffrey–Bolker — LessWrong](#)

Natural Abstractions

- **Basics of Bayes nets/causal DAGs; & causal inference**
- Frankly I don't really know what else needs to go here
 - Category theory might be helpful for the notation and approach?

Elicit latent knowledge

- **Basics of Bayes nets/causal DAGs**
- Mechanism design/decision theory

Norman Hsia

started his own research project after the ai safety camp
trying to find a language with which to describe intelligent behaviour of agents, to talk about safety properties sensibly

started before official start date

early december 2024

abram compiled structured reading list

- started reading that
- important ingredients- studying logical induction - miri came up with
 - logic + probability
- discussion with other mentees

official start of aisc

- weekly group meeting
- personal one on one abram meeting each week

at start, ideas arose through reading - not necessarily research, but more 'what if we do x in y way'

tried to see if could write down useful things about these ideas

put restriction on self 'write 1 page of what to talk about before weekly meeting with abram' - otherwise might ramble

- immature writing, not necessarily leading to something - building a pool of potential starting points that you can -
- heard from guest speaker "spend half of time thinking about what's worth doing"
- 1st concrete
- another mentee noticed/felt flaw in assumption in abram's tiling theory - constructed real world thing where it didn't work
 - new version turned out to work
 - was some back and forth with abram
 - made math v complicated
 - large part of complicated part didn't work
 - but if throw that away, still worked
 - if strengthen another assumption, then the weakening of this one was better, and allowed to remove a lot of the new math complication
 - abram found logical gap
 - with somewhat more reasonable assumption

Cole Wyeth

wanted to extend the utility function of aixi

talked to marcus hutter the inventor (also advisor) of aixi

read some background literature

had to define an integration - not trivial

- two ways
 - end of interactions - no longer predicting
 - standard measure theory

'invented a different type of integration' - 'extended the standard type of integration'

realized it's equivalent to another type of integration under some conditions - wouldn't have been able to tell if hadn't done that previous work

noticed there should be a connection

- did this by rewriting it in a way that reminded him of something that vanessa had done
 - was able to do this because of all the math and integration he'd done before
- was talking to vanessa kosoy
-
- caused him to look into imprecise probability
- spent time looking into computability details

still need to prove that the agent is well defined

even if can define expected utility, not nec that there's thing that maximizes it

- for this, read (had already read) paper that did somewhat similar thing
- - proved existence of aixi with discount factor
 - also generalized aixi utility function, but less so
 - this form of argument(in the paper) could also work, but less so - so work was extending this
 - being able to read and understand the paper required a level of mathematical maturity

wrote the paper

got feedback, responded to it

lot of time was spent on making it accessible

25% of time spread out over 1 year - not possible to actually do in 3 months - need to sit on an idea, take breaks, etc,

Sahil

claim

- current research infrastructure isn't equal to task of capturing age old questions of agency, power, etc
- 'opposite of abstractions'
- 'patterns of the intersection'
- instead of intersection, connections without commonalities
 - e.g. universal language, or speaking own language with translation
 - lot of work is captured in theory normally

- 'captured in a theory prompt' - what he's doing instead
- comments
- adaptive theory
- pluralistic definition

successes

- paper on adaptive parts
 - wants to formalize sometime
 - dual of concept of causal networks
 - example of 'dual' - gcd
 - worked with sophie lipkin and mathew far
 - they were able to come up with definitions that were relevant
 - only got a partial contract, so weren't able to do as much as wanted
- why do this rather than other things?
 - when dealing with highly adaptive things such as ai, are able to track some things that stay fixed
 - fluid theory will be doing that adapting part
 - currently limited by mathematical capability of ai models

Links:

- <https://www.lesswrong.com/s/aMz2JMvgXrLBkq4h3/p/9KamjXbTaQpPnNsxp>
-

Lucius

working on mech interp - thinks its the biggest bottleneck

decomposing nns into circuits

v different from saes, transcoders, etc

Learning theory

Solomonoff induction - instead of doing full si over all program, we just do si over all program that are 1gb length

- also, not gonna give infinite comp resource
 - not count if not fin in some no steps
 - now we have computable algo - finite program, finite
 - uniform prior over these programs
 - do bayesian updating over incoming data

take space of all programs

2 programs that could fir

1 is 1 kb, other is

Vanessa Kosoy

Missing emphasis on theory

every problem in here:

<https://www.lesswrong.com/posts/ZwshvqiqCvXPszEct/the-learning-theoretic-agenda-status-2023>

Various aspects of learning theory we don't understand well

If learning theoretic agenda was solved, then what:

- propose algorithms/architectures for ai systems that we can prove solve alignment

Estimate of how long for the agenda to be solved (rough estimate):

- 40 researcher years - very rough

Progress in the last 2 years

- 1.5% roughly

Main bottleneck:

- number of people working on it
- number of people available
- number of people Vanessa can teach

Who else can be a mentor:

- Alex Appel - Diffraction

If had math phds for 5 weeks

- write up something that has basic stuff written in it
- 5 weeks is short
- says at least 4 months is minimum for making progress
- if computational learning theory
 - able to dive in more quickly
 - even better if theory of rl
 - e.g. proving regret bounds

Is there something that's useful to implement yet in LTA?

- early
- maybe some specific experiments
- dl algos in adversarial conditions
- not really the area where she thinks useful to look at

Most useful thing if they're a quant/competitive mathematician

- learn the math in

<https://www.lesswrong.com/posts/ZwshvqiqCvXPszEct/the-learning-theoretic-agenda-status-2023>

Measure of success of program:

- if someone writes alignmentforum post that Vanessa considers useful
- 5 weeks is short
- if someone writes post/work afterwards and says the program was what got them started
- Posts Vanessa found useful/insightful
 - https://www.alignmentforum.org/users/abramdemski?from=search_autocomplete

- <https://www.lesswrong.com/s/mqwA5FcL6SrHEQzox>

Vanessa's research process:

- backchaining process
- what concepts are we concerned with?
- what theories that tell us something about the concept?
- what are the theories missing?
 - e.g. assumptions that are problematic
- what's the simplest mathematical toy model that could address this?
- simple but not too simple - actually contain a hard part of the problem - not vacuous
- try to mathematically formalize it, prove it
- what have i learnt from this?
- maybe proof taught me i was looking at the wrong framework
- if successful, maybe can expand to more complex, reduce some assumptions, make less toy of a problem
- maybe brought up new questions that didnt see before seeing the solution

What part could be automated?

- the part where we have a formal mathematical problem and need to prove it - *maybe* it could be, not sure if the tech is there yet
- currently finding it most useful for literature search
- when actually have question Vanessa doesn't know answer to - not good enough, hallucinated answer thats wrong upon inspect - havent looked very hard

Is there a part of the process that can be handed off?

- not sure
- maybe the writing up - can be very time consuming - but to do it, would need to understand it
- would be nice to have pedagogical materials
 - textbook style

When onboarding is there a particular part that's faster/slower to teach people?

- lots of ideas Vanessa has that's not written down neatly
- <https://www.alignmentforum.org/posts/wzCtwYtojMabyEg2L/what-is-inadequate-about-bayesianism-for-ai-alignment> Britnay has written some stuff
- learning theory course is recorded and notes will be uploaded

Needed for Learning Theoretic Agenda:

- Decision theory can be picked up
- important parts:
 - theoretical computer science
 - complexity theory
 - probability theory
 - advanced
 - e.g. measure theory,
 - Most useful:
 - statistical learning theory
 - computational learning theory
- useful:

- category theory

In 5 weeks with Computational learning theory:

- <https://www.lesswrong.com/posts/m4NMk6EinRzvvvW5Y/closed-gauging-interest-for-a-learning-theoretic-agenda>

In team:

- bottlenecked on learning on what the problems are
- background
- if 2 or 3 months, then can make substantial progress if working on a very well defined problem
- finding the problem takes a long time - very roughly - usually at least 2 weeks - to wrap their heads around the problem

Specific Problems/**Bottlenecks**

if someone has computational learning theory background - imprecise linear bandits, maybe solve it within a month,

if category theory, post on string machines, work out details more rigorously,

category theory, shortform on string diagram "diagrammatic notation for constructing new infrakernels out of given infrakernels. There is probably some natural category-theoretic",

> when you have a kernel, that's a function that takes a state and action produces a probability distribution over the next state of the system infrakernel except instead it returns a set of probability distributions

shortform on "Systems which locally maximal *influence* can be described as VNM decision-makers."

- general Probability theory background
- topology functional analysis

bar for promising researcher:

- theoretical computer science researcher - published work

Roman Malov

Part of Abram's group

Didn't make progress

No substantial progress on his part

Bayesian Logical Inductor

trying to provide a constructive way to have a belief system about its future beliefs

- logical inductor changes its beliefs
 - thinks, in some sense, then changes its beliefs
 - changes its beliefs day by day
- bayesian logical inductor changes its beliefs

- observes some events
- based on some events it changes its beliefs
- tried to merge those two things
 - tried to provide a probability distribution by observing in which states are logical inductor
 - bayesian logical inductor looks at logical inductor on some day
 - has beliefs on mathematical statements
 - trusts logical inductor if it thinks on something for a long time
 - a way to update for small beliefs
 - logical inductor, when it thinks for longer, it considers more statements
- was trying to figure out what would be the prior distribution on large beliefs
- somewhat arbitrary - was trying to find elegant way to do it
 - interesting in some point of view
 - to be consequential, need to make logical induction feasible and comparable to other ways of reaching mathematical truth
 - use updateless decision theory
 - uses bayesian updates
 - needed to some
 - hope is logical induction is in some way possible to use it to construct a general engine
 - hope is not just mathematics, but other domains as well
 - e.g. axioms about other domain
 - need to make logical induction consider not just mathematics but other domains and incorporate
 - make logical induction faster
 - the current github:
 - <https://github.com/monasticacademy/logical-induction/blob/master/logicalinduction/sentence.py#L12>
 - relies heavily on traders
 - in LI, take every possible program, see which every outputs strategy
 - obv way, take not every one, but some reasonable amount, which are somewhat reasonable traders
 - maybe turning deepseek prover into a trader?
 - maybe a bunch of traders?
 - logical induction can be thought of as scaffolding for a bunch of minds to find truths in mathematics
 - Roman's Idea: you can plug in traders who are going to have reasonable guesses about what mathematics is going to look like
 - not been done yet
 - Roman may try it

in some sense made progress, but didn't feel like it was useful -

Can agent foundations be implemented?

- maybe, but wouldn't be very impressive

- could do it in 1966, but you said more words about it
- one of the ideas of Abram is related to LLMs
 - dream true and good machine
 - how to make llms more truthful
 - separate the levels of training to be good, true and imaginative
 - point of logical induction is to have fast probabilistic estimates of mathematical statements
 - doing this in a systematic way

How pick which problem to work on:

- in some sense didn't
- Didn't think AI Safety via debate etc would scale
- Agent Foundations didn't seem to obviously not scale
- was more familiar with Abrams work

For someone with theoretical computer science bg

- level above

Roman's background:

- bachelors in theoretical and general applied physics
-

What he read in order to get started:


stuff from here:

https://docs.google.com/document/d/1JG51VLMMkMugxU3qI_ZFudj2Aj4jnJqiv71Lf84DOi0/edit?tab=t.0#heading=h.9lmc73wscx1r

what was necessary:

- Understanding Trust
 - what learnt:
 - sketch proofs
 - the rough proof - the version of the proof that needs refinement to become formal
 - what is udt
 - the specific version
 - diff between udt 1.0 and 1.1
 - actions, policies, agent tries to reach optim policy, what that means, how it estimates that, some actions are self modifying, some are not, what specific assumptions we make about the real world, the more reasonable the assumptions the better, if proofs are actually making programs
- Logical Induction
- Maybe read something about Updateless Decision Theory
- Maybe read something about Tiling

-

-  Understanding Trust - Readings - read 30 - 40% of this and was basically fine
- Should be more than sufficient - dropbox lectures

progress of maths is if you cant prove it, if you can't backtrack from it to

Cameron Holmes

Picking a good project

- sharing with peers

Case for alignment projects from ukasi

<https://www.lesswrong.com/posts/iELyAqizJkizBQbfr/an-alignment-safety-case-sketch-based-on-debate>

<https://www.alignmentforum.org/posts/hjMy4ZxS5ogA9cTYK/how-i-think-about-my-research-process-explore-understand>

MATS teams - 8 researchers in a group, sharing ideas, find things that naturally cluster, find appropriate teams - mostly organic, not trying to prescribe - most important thing is that they're excited to work together

come up with cheap experiments to test ideas

Comments on structure: - timeline is compressed - shift timeline from being a Result to a Plan - to be a Scholar Research Plan

MATS program - 10 weeks

MATS extension - 6-10 months

Set expectations

Scholar Research Plan

- solid theory of change, have some impact, solve part of alignment problem - 60% of importance
- resources needed
- milestones - high level
- elaboration around thoughts for the project

What they need to have done to make a good SRP

- red teaming ideas
 - 90% of the feedback from peers - but the final thing coming from senior researcher
- small experiments

- just a bit of work to turn this into a grant application (not a part of the plan)

Most important thing is coming up with a solid SRP - make it clear that a Poster or Paper is not the needed thing - best thing is a really solid plan.

What makes very talented people be attracted:

- Mentors - very high class mentors
- Previous expectations of outcomes

Thinks companies won't pay for job booth, because they're starved for headcount. Orgs don't have that much headcount to hire from - doesn't hurt to ask.

Want to know before recommended - who mentors, how much time

Abram Demski

skillset

- Mathematical philosophy
- philosophy skills
 - looking for counterexamples to your idea
 - grappling with informal ideas and trying to make them more formal
 - analyzing arguments
 - basic critical thinking skills
- mathematical skills
 - formal proof skills
 - strong mathematical bg
 - different aces that might be useful
 - able to make own math - related to formalizing the informal
 - able to read math
 - Definitely:
 - probability theory
 - decision theory
 - expected utility theory
 - logic
 - theory of computation
 - representation theorems - Savage - first to justify EU and prob theory together, Jeffrey, von Neumann theory, Dutch book arguments
 - radical probabilism
 - **Good to have**
 - algorithmic information theory
 - some kind of learning theory is important - for where the problem needs to head - not sure what to recommend

-

Specific Research bottlenecks

- integrating theory of logic uncertainty or computational uncertainty with udt setting
 - lot of complexity of logical induction algo and tools needed to integrate that with udt
 - gets the sense that it *might* be easier to invent a different theory for logical uncertainty, specifically to fit updateless decision theory
 - still compatible with radical probabilism
 - figuring out notions of coherence that go beyond the logical induction criterion
 - specifically in relation to large sentences
- theory of ontological uncertainty
 - how can one agent trust another agent when they don't share the same probability/sigma algebra
 - event space underlying a prob distribution,
 - the same way of looking at the world
 - how they parse everything
 - both the underlying senses could be different (vision vs hearing) and the way they group those into objects could be different
 - theories of natural abstraction
 - e.g. wentworth
 - sam eisenstat (not sure if he's published anything - but should publish soon)
 - the goal is to provide a justification for a particular theory, through this trust theorem
 - alice looks at bob
 - alice has some degree of trust that even if alice sees basketball and has particular thoughts as a result of seeing it, bob sees it as a set of atoms, still alice has some sort of trust with bob
- handling value change while keeping trust
 - radical probabilist ideas v relevant to this, provide pic of belief change that readily adapted to pic of value change
- robustness under value uncertainty
 - things like bargaining theories
 - bunch of stakeholders at table, don't want any particular stakeholder to lose out too much
 - nash bargaining solution
 - basically just x all utilities together rather than weighted sum

- provides pret good theoretical guarantees,
 - but doesn't work if one stakeholder has exactly negative utility of another
 - - want ai system - has uncertainty of what human values are, has hypothesis of what they are, has semi possible thing, we want to be sure that our values are in there and that any values in this domain has a pretty good chance,
 - - we're starting with v broad uncertainty about what human values are
 - definitely things in there that are impossible to bargain with
 - - e.g. flipped utility
 - Nash Bargaining seems sort of close to what we want but won't actually work in practice
- cooperation and coordination
 - coop means prisoners dilemma
 - coordination - stag hunt
 - problem to solve
 - this eq to the whole problem when viewed in right way
 - thing to do is create theorems in which coordination and cooperation to happen
 - want increasingly finegrained and useful pictures of how humans relate to ai
 - both positive theorems and negative theorems
 - these are pictures of general agents relating to each other
 - if want to understand beneficial relationship to ai - then want to understand beneficial relationships in general
 - old school thing 'if we want friendly ai, we want theory of friendship'

Most doable for someone new to agent foundations specifically

- someone with bg in theoretical computer science
 - robustness under value uncertainty, most useful to work
 - after that, all of them somewhat tractable,
 - least, integrating theory of logic uncertainty or computational uncertainty with udt setting
 - mainly because it seems hard
- could someone identify for themselves if they can work on the problems
 - each can be tackled in way not helpful

- keep goal in mind - make ai do good things not bad things
- keep trying to, as clearly as possible, how it plugs into bigger picture, how work is going to create positive visions of future

Not excited about empirical prompt

- some experiments interested in, but not for this list of things

Thinking through ideas

- how identified above as bottleneck
 - top goal - ai good not bad
 - some picture
 - keep refining
 - some picture you've made, not someone else has made, so you're running it through your cog process, continuing to refine that in response to reading other people's stuff, in response to what seems right/wrong about it
- yes, working on problem above
- how thinks through problems above
 - come up with simplest picture that still has pic of thing
 - important to first do simplest example you do know how to do - wasted lot of time *not* doing that
 - accidentally cut off thing he didn't know how to do
 - do that
 - keep complexify that
 - what's the next most important thing want to capture
 - start with very simple tiling proof
 - keep saying what's wrong with it, what improve
 - when sitting down with specific chunk bitten off
 - need a proof idea
 - intuition of how proof is going to go
 - talking to yourself
 - this guy can trust this guy because
 - sometimes easier to write down assumptions that are prob needed
 - then stick those into proof
 - sometimes easier to write proofs not in math but in english, then figure out which assumptions needed to justify each step
 - then write math for the english, prob something wrong, fix until working math
 - lot of basic skills in this domain, unpacking definitions, lot of skills you get when had proof experience
 - diff thing compared to pure math, is deeper informal intuitions due to being human

- very important to generally take notes
 - crucial to be writing a lot
 - capturing ideas somewhere, important in telling brain to generate more of those ideas, helps see the whole thing at a glance, see how it fits together/not fit together
 - lets you refine it in way diff than if just in your head
 - and diff than talking to people - that important too, but different
 -
 -

the notes should be in math and natural language, switching back and forth

If had math phds for 5 weeks, what have them do

- run same program that ran in spring [Understanding Trust - Readings](#)

Why not particularly excited about the applied approaches

- undertaintly about how to turn emphical uncertainty into i
- not dramatic emphircal is useless
- not sure how to generalize known experiments into something that will scale to superintelligence
- is it worth to test which experiments *don't* scale?
 - does that test itself scale to superintelligence?
 - value in ruling things out?
 - could be interesting to see if an experiment inspired by the theorem scales/not
 - known experiments that have been tried to be inspired by agent foundations
 - closest was rlhf
 - going by clustering of people - christiano
 - model inspired by thinking about the problem in agent foundationy ways

What would good output of 5 weeks look like?

- would love to see
 - as example - ans to question - what
 - conception frame has
 - kind of true, not
 - bayesianism, first fully formal paradigm for science, give one baby pic of trustworthy reasoning
 - too many people in science for this to work
 - frequentism works as theorm for individual science, trying to advance science together
 - scale science even more, then too many scientists, they discover p hacking, not robust anymore
 - people have new ideas for how to make things adversarial ideas
 - llms then break everything

- what's the new science that lets you take IImms and have useful science
 - when you try to get truth out of them, you get hallucinations
 - what's the new scientific paradigm for the new cognitive paradigm
 - might look like some kind of market paradigm
 - might look like logical induction, but not logical induction, since logical induction is not feasible to run
 - theory of communication among semi cooperative agents
 - coordination successfully most of the time
 - theory of what they comm to each other to maintain coop with each other, looks like market, maybe virtually betting on each statement/claim, when making request - bid for this to happen, some amount of money on that, some sort of robustness to it
 - can easily generalize value and
 - if all markets
 - value change is change in markets
 - all fluctuating all the time anyways
- don't know how to translate to geometric rationality type stuff
 - money is linear
 - market based thing that recognizes that agents in market have goals
 - want robustness against things like company taking over country (if modelling country as market)
- formalization of a market that's robust to value change, exploitation, hostile takeover, very large number of market participants
 - robust (in the sense that gives trustworthy predictions and recommendations)
 - theorem/theorems shows if market is robust/trustworthy
- useful negative results
 - showing that we *can't* establish this kind of justified trust with certain conditions
 - new articulation of failure modes
 - things that could lead to existential risk
 - things that could lead to a great deal of suffering
 - e.g. inner alignment vs outer alignment was a big new thing at one point
 - prev, ai safety people had just been thinking about outer alignment
 - comes from getting a clear picture of some plan, that hasn't been articulated this way before
 - one way it can happen - maybe can happen in other ways as well

What would a deceptive bad output be? something that they think is good, but not

- Literature review for cooperation in game theory that's already known
 - shadow of the future
 - iterated prisoners dilemma
- Getting *too* anchored in how humans solve problems here
 - humans can give inspiration
 - quite skeptical we've solved how to trust superintelligence
 - e.g. 'good parenting skills' - not scaling to aligning superintelligence
- sure there's a lot of ways to go wrong here

Who else could be a good mentor?

- maybe think of them later

Could any parts of this be automated?

- have had useful conversations with ai, but not automated at all, ai will actively try to trick you with fake proofs that look good at first glance and you really have to check through
- feels too conceptual for automated proof search

On mentoring - Between **August 2nd and August 16th**

Talk - August 3rd

Resources that are helpful:

- general notetaking

Scott Garabant - not a good time to bring new people

Tentatively yes, for recc for donation

People trusted:

Scott Garabat, Vnessa kosoy, Sam eisentat, paul christiano, nate, eliezer, samuel buteau

- not trusted in everything
 - e.g. christiano working on heuristic argument that's completely doomed
 - most people think most things doomed, doing least doomed agenda
- trust about as much as own judgement

tsvi

context: not currently pursuing alignment research, pivoted away in 2024, was already pivoting away, think it's too hard

unlikely to be leading research even in his type of research
people would need to self direct or find other

previous mentee: mateusz baginski

What was research: solve hard part of alignment problem

hard part of alignment problem:

- understanding enough to design a mind that's genuinely smarter than humans but does stuff that we like and doesn't kill everyone
- not 'human values'
 - need corrigibility
 - naive way of 'human values' not compatible
 - problem of [fully updated deference](#) - prob by eliezer, on arbitral

The problem of 'fully updated deference' is an obstacle to using [moral uncertainty](#) to create [corrigibility](#).

One possible scheme in [AI alignment](#) is to give the AI a state of [moral uncertainty](#) implying that we know more than the AI does about its own utility function, as the AI's [meta-utility function](#) defines its [ideal target](#). Then we could tell the AI, "You should let us [shut you down](#) because we know something about your [ideal target](#) that you don't, and we estimate that we can optimize your ideal target better without you."

The obstacle to this scheme is that belief states of this type also tend to imply that an even better option for the AI would be to learn its ideal target by observing us. Then, having 'fully updated', the AI would have no further reason to 'defer' to us, and could proceed to directly optimize its ideal target.

Furthermore, if the present AI foresees the possibility of fully updating later, the current AI may evaluate that it is better to avoid being shut down now so that the AI can directly optimize its ideal target later, after updating. Thus the prospect of future updating is a reason to behave [incorrigibly](#) in the present.

While moral uncertainty seems to take us [conceptually](#) closer to deference-based [corrigibility](#), and there may be research avenues for fixing the issue (see below), the current explicit proposals will (when scaled to sufficiently high intelligence) yield essentially the same form of incorrigibility as an AI given a constant utility function.

-
- prev way of approaching this was 'value uncertainty'
- if does autopsy
- worries are not just "wrong utility function"
 - impoverished view

Research:

- trying to understand minds
- what ends up determining
- if you have a mind, that's very capable, there's a free parameter of what it makes the world look like
- 'mostly not working on specific problems'
- 'might not like this, fair to not like this, but will stand for this'
- thinks we have to take a very circumspect approach, not bake in assumptions about specific
- we're deeply confused about what determines what a mind will do with the univers
- pay attention we're deeply confused about that

What working on alignment research looked like:

- doing a bit of math
 - reading lesswrong/elizier's work, timeless decision theory, creating friendly ai
 - prob theory, logical uncertainty, decision theory
 - what problems trying to solve with the math?
 - for decision theory, looking at toy problems, where behaviour interesting
 - reflective stability
 - the idea of trying to understand when it's the case a mind would want to self modify in a deep way
 - one way of exploring that is (mentions newcombs problem and asks if I've heard of counterfactual mugging)
 - hard to
 - 'if you make decisions in this way in general, you'll do better/worse'
 -
 - asymptotic decision theory
 - a decade ago
 - Motivation:
 - when you look at counterfactual mugging
 - it's a toy model, tells us something about 'reflective stability'
 - reflective stability needs to be understood or you're completely lost
 - if you don't understand when a feature is stable
 - what feature could determine what long term effect a mind has on the universe
 - if that feature is not
 - if you can identify the
 - instead of looking at one problem
 - don't think he had that much research success
 - helped writing logical induction, but that was mainly scott's work
 - what trying to specifically find out and how
- 2015 - 2019
- all about reflective stability
 - using toy models
 - proof search is a general mathematical searcher
 - what problems
 - looking at various papers,
 - counterfactual mugging, open source decision theory, agent simulates predictor

- most important - in retrospect, not important, because the wrong problems to look at,
 - came to think
- 2019 to 2023/2024
- 'current ais are not real minds'
- 'our access to animal minds isn't very good'
- 'our brain scans suck'
 - not getting very high resolution in space or time
 - can get high resolution in space but not enough data
 - not sure if that's even a crux
 - even if very good scans and paired with ethical morality
 - if had an ethical dilemma for several months/weeks
- more philosophy of mind approach, still reflective stability
 - didn't meet much success
 - pointing at directions, point at obstacles, etc
 - tried to read about ethology
 - What would you do during a day?
 - walking around and thinking
 - little uncommon thing that might be important
 - meditative kind of way - abstract, high level, meditating, looking (had tangent about the gita and upanishads again)
 -
 - talking to collaborators
 - with mentees, have google doc open, then ask what's been on their minds, give idea/question, they'll ask something he said,
 - have conversation
 - he'll be like 'here's an important thing' and type it into a doc
 - mostly stopped trying to directly collab on research at like 2019/2020ish
 - before spent lots of hours arguing about what kind of thing is needed for alignment
 - conversations with collaborators with people like sam eisenstat
 - mentorship
 - trying to write things
 - discuss things with mentees
 - if things that i find myself arguing a lot
 - if i have opinion that seems important
 - thing that's good/not good
- trying to think
- formal analysis,

- process of identifying if a research idea is going to be good
 - has changed
 - 3 phases
 - 2019
 - is this on the path to answering technical question someone else has posed
 - 2021
 - is this/does this seem like a novel approach to getting contact with what matters about minds, reading about neuroscience, evolution, looking at evolution as an optimization process
 - approach here: is this a good way of getting into contact with minds, animal behaviour, ethology, would have looked at interpretability, but that was only starting up
 - 2021-2024
 - not is this solving the problem, but is it supporting the mental operation of fixing mental gaze more intently on the actual problem of alignment or the actual core problem or subproblems
 - multiple bottlenecks
 -

human intelligence enhancement

- want to increase problem solving ability
- lots of important things, such as getting lead out of pipes, domanine, etc
- technological approach
- <https://tsvibt.blogspot.com/2024/10/overview-of-strong-human-intelligence.html>
-

Sam Eisenstat

agent foundations

find shared meaning that communities of agents can find true

different from natural abstractions

- disagree with both the word natural and abstractions

2 versions of theory -

- strings
- random variables

we have shared world and we describe in terms of certain latent variables

in johns we dont have lots of variables

in johns there is a bit objectivity

it's like sufficient statistics

variant equiv

latests share prob best statistic

causal model where many different latent variables

want structured world with many different variables

want quantificational logic

constructing family of latent variables

want to destroy supervised rl or reinforcement learning as being the only way of doing things

current research bottlenecks

depends how broadly speaking

want to validate framework

- bunch of directions its validated in

develop it futer

relation to truth and quatatil logic

using condensation to tell sotries about vaes/learnig/agents engaged in coolb actions where problem in meaning, figure out concepts/shared concept, how to explain things to each other

inner alignment stuff, apply to neural nets

- some success in interp paradigm, find some variables which meaningful with human judgement

what working on rn:

- developing and validating framework
-
- want to apply to statistical framework

main advantage between this and neural networks

- can't really say it's stocastic model because in some sense nn is also stocastic

- not have theoretical tools to analyze two different models (neural networks)
- would be nice if could be

Talk

- context of the research
- the mathematical ideas
 - random variable models and latent variable models
 -

Given probability distribution

model/analogy:

i have a coin, i flip the coin 10 times, i'm given joint distribution of 10 coin flips, the shared thing is the bias of the coin, if have y

ml is missing a lot of agent phenomena - want to reconstruct everything on a very different basis

look for the concepts in his theory, look for the hidden variables

What new thing are we learning from this theory

- not just distribution on observation sequences, but on sequences??

when we understand an agent, we should understand what is its world

<https://www.sameisenstat.net/doc/condensation-25-07.pdf>

take some interesting joint distribution and figure out what are the latent variables, how much room do we have to move latent variables around, what are the wave packets that move around like particles,

- this is what currently working one

Neuroscience Researchers

Nich

insular cortex - 'plays a big role'
'legioning studies'

yacine (neuroscience, phd in consciousness)

imaging element of the brain in flux
test for consciousness - knock the person out with propofol

- see it come back
- most changes are most likely to come back

Chloe Loewith

fmri imaging is best we have and v noise

- could reach out to sister
- denoising the data
 - need to talk to the clinical labs

start in philosophy

brain

marmasets speak to each other, what light up - in fmri image, what regions active

main question development - philosophers to start the big questions

- analytical phil
- bioethics
- how to frame
- whats the timeline
- what exact parameters
- strict diameters
 - example of strict parameters
 - she's in ai ethics genemoics
 - define timelines
 - neurotechnology can mean anything - specify definitions
 - miscommunication can happen at any level - make sure everyone knows what they

Luiza

was in spar

- value alignment in collective intelligence systems
- started with whats alignment, whos alignment do you align to
- what is collective intelligence systems
- limited to language models without tool usage
- to the core was a round robin communication
- what did:
- decentralized and centralized thing
- moe
- looked at diff performances
- for value alignment
- values in the wild
- analyzed subset of convos with claude and labelled them
- took some of the values they had
- used similar subset of their methodology
- tested on other models
- used language models as a judge, to label transcripts based on these values
- was first time, was language
- didnt have enough time, 3 different continents
- lost track of initial goals
- help people be really focused with whats the deliverable at the end
- did lots and lots of ideas, had lots and lots of ideas

did they have intermediate deadlines?

- tried to have weekly checkins with mentors
- lasted 2 weeks
- having regular checkins would be good
- rather than just async

Make the feedbacks to teams be public and viewable to everyone

- all meetings were like: "what are the tasks?"
- everyone wanted to be told what to do
- is the thing of what to be done the actual thing
- good management and planning needed
- **lacking Clear Deliverable**
- How team decided what idea to start with:
- in February, there was lots of very cool papers coming out
- In SPAR, the *mentor* came up with the idea
- lots of ideas
- scope was very big
- team was very international and had different values, so 'which values' was

Thoughts on Demo Day

- hesitant voice: was nice?
- nervous to talk to more than 3 people
- practice session for presentation evening: Could be useful
 - friends **always** have a rehearsal

After the program:

- not success in this
- mentor got very busy
- their team discord isnt really active
- still in touch with their teammates

What asked researcher:

- wanted to know how research is done
- mentors were also figuring it out themselves
 - platitudes
 - lots of talk on possibilities, not actual experimental design
 - what will be the actual code
- not that the doc was bad, but not all of it was essential, didn't say what is important and why
 - were also changing version
 - some updated, some not
 - big consistency problem

What others experience was like;

- SPAR also organized coffee chat
 - the people organized with, didn't show up
 - weren't asked 'do you want to participate, do you not want to participate'

Would you take part in SPAR again?

- yes
- spending time with other, technical people is good
- the problem is still intersting
- would totally recommend it
- not the next session - hopefully doing something better

Anything else?

- Having clear deliverable - goal of publishing something, with hope of further expansion
 - extremely well defined target

interpretability at AMD

intersection of neuroscience and interpretability

in the process of learning

in choosing research ideas - falsifiability

Sahba Afsharnia

denoise in ig comp rl

Neuroscience Experts

Subject: Moral Encoding in the Brain - Research bottlenecks

Dear __,

I'm Kabir Kumar, I'm running an alignment research team. I am reaching out to you because of your work in understanding how morals are represented in the human brain.

We're currently organizing a [5 week research program](#) focused on building and improving AI architectures that encode values in a way that generalizes robustly, scales gracefully, and reflects meaningful internal representations, rather than surface-level behavioural patterns or shallow refusal mechanisms.

As part of this effort, we're speaking with neuroscience researchers like yourself to better understand the current landscape and learn what are the most useful problems for researchers to work on. I want to know: What are the most significant bottlenecks you're facing when it comes to mapping value representations in the brain?

This is both to make a guide for participants on how to do valuable research during the program and to know how to best structure the program so that the most important problems get targeted.

Would you be open to a short call sometime in the coming weeks?

<https://calendly.com/kabir03999/talk-with-kabir?month=2025-06>

Looking forward to hearing from you,

AI Plans

NeuroScientists to Interview

Neuroscientists we'd like to interview: anyone who's worked on finding out how the brain encodes values/preferences:

. Foundational Pillars & Neuroeconomics Pioneers

These are the figures who built the bridge between economics, psychology, and neuroscience, establishing the field's core questions and methods.

1. **Paul Glimcher** (New York University)
 - **Contribution:** Considered the "father of neuroeconomics." Pioneered primate electrophysiology to show that neurons encode subjective economic value, not just physical properties of stimuli. Author of foundational texts like *Decisions, Uncertainty, and the Brain*.
- 2.
3. **Antonio Rangel** (Caltech)
 - **Contribution:** A leader in human decision neuroscience using fMRI. Champion of the "common currency" hypothesis, proposing the ventromedial prefrontal cortex (vmPFC) as the brain's central valuation hub. Developed the attentional drift-diffusion model (aDDM) linking attention and choice.
- 4.
5. **John O'Doherty** (Caltech)
 - **Contribution:** A master of using computational reinforcement learning (RL) models to analyze fMRI data. His lab produced seminal work identifying reward prediction error signals in the human striatum and dissociating different valuation systems.
- 6.
7. **Wolfram Schultz** (University of Cambridge)
 - **Contribution:** A Nobel Prize-winning neurophysiologist who discovered that dopamine neurons in the primate brain broadcast a reward prediction error signal, a finding that provided the biological basis for modern RL theories of learning and value.
- 8.
9. **Read Montague** (Virginia Tech / UCL)
 - **Contribution:** A pioneer in computational psychiatry and human neuroimaging. His work was among the first to demonstrate reward prediction error signals in the human brain (using fMRI) and has been crucial in applying these concepts to understand addiction and mental illness.
- 10.
11. **Peter Dayan** (Max Planck Institute for Biological Cybernetics, Germany)

- **Contribution:** A leading theorist in computational neuroscience. His work on reinforcement learning, dopamine, and Bayesian models provides the mathematical and conceptual foundation for much of the experimental work in the field.
- 12.
13. **Alan G. Sanfey** (Radboud University, Netherlands)
- **Contribution:** Conducted the landmark fMRI study using the "Ultimatum Game," which launched the field of social neuroeconomics by showing how brain regions involved in emotion (insula) and cognition (DLPFC) interact to process fairness and value.
- 14.

II. Core Valuation, Choice, & Computational Modeling

These researchers are at the heart of the field, dissecting the algorithms and neural circuits for learning and representing value.

A. Human Neuroimaging & Computational Modeling

1. **Nathaniel Daw** (Princeton University)
 - **Contribution:** A leading figure in applying RL models to neuroscience. Famous for his theoretical and experimental work distinguishing "model-free" (habitual) from "model-based" (goal-directed) valuation systems in the brain.
- 2.
3. **Yael Niv** (Princeton University)
 - **Contribution:** A major force in computational neuroscience, focusing on how attention, memory, and latent state inference shape learning and decision-making. Her work connects reinforcement learning to broader cognitive processes.
- 4.
5. **Joseph Kable** (University of Pennsylvania)
 - **Contribution:** Known for meticulous and influential fMRI studies on subjective value, particularly in the domains of intertemporal choice (patience) and risk aversion. His work provides a clear neural signature for subjective value discounting.
- 6.
7. **Brian Knutson** (Stanford University)
 - **Contribution:** Pioneer of the "anticipatory affect" model. His work using fMRI clearly dissociates the role of the nucleus accumbens in anticipating potential rewards from the role of the insula in anticipating losses.
- 8.

9. **Ray Dolan** (University College London)

- **Contribution:** A giant in affective neuroscience and computational psychiatry. His lab has made profound contributions to understanding how emotion, learning, memory, and neuromodulators (like dopamine and serotonin) shape decision-making in health and disease.

10.

11. **Peter Bossaerts** (University of Cambridge / University of Melbourne)

- **Contribution:** A unique researcher bridging neuroeconomics with computational finance. He studies how brains handle risk and uncertainty in market-like settings, testing formal economic theories with neural data.

12.

13. **Valerie Reyna** (Cornell University)

- **Contribution:** Proponent of "fuzzy-trace theory," a dual-process model suggesting that decisions are often based on gist-based, categorical representations ("some vs. none") rather than precise value computations.

14.

15. **Hackjin Kim** (Korea University)

- **Contribution:** A prominent researcher in affective and decision neuroscience, known for work on the neural basis of anxiety, how it affects value-based choice, and the role of the amygdala and prefrontal cortex in processing uncertainty.

16.

17. **Robb Rutledge** (Yale University)

- **Contribution:** Develops computational models of subjective feelings, particularly happiness. His work links momentary happiness to expectations and reward prediction errors, creating a quantitative bridge between valuation and emotional state.

18.

B. Primate & Rodent Neurophysiology (Cellular & Circuit Mechanisms)

1. **Daeyeol Lee** (Johns Hopkins University)

- **Contribution:** A leader in primate neurophysiology, studying the role of the prefrontal cortex and basal ganglia in complex decision-making, reinforcement learning, and strategic thinking.

2.

3. **Camillo Padoa-Schioppa** (Washington University in St. Louis)

- **Contribution:** Has produced exceptionally clear and influential work from single-neuron recordings in monkeys, showing that neurons in the orbitofrontal cortex

(OFC) encode the subjective value of goods in a "common currency" format, independent of motor action.

4.

5. **Benjamin Hayden** (University of Minnesota)

- **Contribution:** Studies the neural basis of cognitive and economic decisions in primates, with a focus on foraging, information seeking (curiosity), and the role of the anterior cingulate cortex (ACC) in monitoring and guiding choices.

6.

7. **Michael Shadlen** (Columbia University)

- **Contribution:** A foundational figure in decision neuroscience. While focused on perceptual decisions, his work on the drift-diffusion model (DDM) as a mechanism for evidence accumulation has been widely adopted as a framework for understanding value-based choice.

8.

9. **Geoffrey Schoenbaum** (NIDA/NIH)

- **Contribution:** A leader in rodent neuroscience focused on the orbitofrontal cortex (OFC). His work suggests the OFC's primary role is not to store values, but to build a "cognitive map" of the task structure, allowing for flexible, model-based behavior.

10.

11. **Naoshige Uchida** (Harvard University)

- **Contribution:** Uses sophisticated techniques in rodents to study the role of dopamine neurons in reinforcement learning, confidence, and sensory uncertainty, providing a cellular-level view of value-related computations.

12.

III. Self-Control, Intertemporal Choice & Effort

These researchers focus on how value signals are modulated by cognitive control, time, and the cost of effort.

1. **Todd Hare** (University of Zurich)

- **Contribution:** A leading researcher on the neuroscience of self-control, particularly in dietary choice. His work proposes an influential model where the vmPFC computes value, and the dorsolateral prefrontal cortex (DLPFC) modulates this signal to enable self-controlled choices.

2.

3. **Christian Ruff** (University of Zurich)

- **Contribution:** A key innovator in using non-invasive brain stimulation (TMS/tACS) to move beyond correlation to causality. His work demonstrates the causal role of specific prefrontal regions in self-control and social decision-making.
- 4.
- 5. **Samuel McClure** (Arizona State University)
 - **Contribution:** Conducted a seminal fMRI study on intertemporal choice, proposing a dual-system model where limbic regions are associated with immediate rewards and prefrontal regions with patient, long-term choices.
- 6.
- 7. **Roshan Cools** (Radboud University, Netherlands)
 - **Contribution:** An expert on the role of dopamine and the striatum in cognitive flexibility—our ability to update goals and switch between tasks. This is critical for adapting behavior when values change.
- 8.
- 9. **Matthew Botvinick** (DeepMind / Princeton)
 - **Contribution:** A leader in connecting neuroscience with artificial intelligence. His work on hierarchical reinforcement learning provides a powerful framework for understanding how we break down complex, long-term goals into manageable sub-goals, each with its own value.
- 10.

IV. Social & Moral Valuation

This frontier explores how the brain's valuation systems are adapted to handle decisions involving others.

- 1. **Molly Crockett** (Princeton University)
 - **Contribution:** A prominent voice in moral neuroscience. She uses a powerful combination of computational modeling, pharmacology, and fMRI to investigate moral judgments, altruism, the value of punishment, and moral reputation.
- 2.
- 3. **Ernst Fehr** (University of Zurich)
 - **Contribution:** An economist who has been a central driver of neuroeconomics. He collaborates extensively with neuroscientists to study the neural basis of social preferences like fairness, trust, and reciprocity.
- 4.
- 5. **Cendri Hutcherson** (University of Toronto)

- **Contribution:** A leader in building precise computational models of social decision-making. Her research investigates how we compute the value of outcomes for others (generosity) and integrate those "social values" with our own.
- 6.
- 7. **Tania Singer** (Social Neuroscience Lab)
 - **Contribution:** A pioneer in the neuroscience of empathy and compassion. Her work differentiates the neural circuits for empathic distress (feeling others' pain) versus compassion (a warm, prosocial motivation), showing how training can change these value-laden responses.
- 8.
- 9. **Rebecca Saxe** (MIT)
 - **Contribution:** The world's leading expert on the neural basis of "Theory of Mind" (thinking about others' thoughts), centered on the temporoparietal junction (TPJ). This capacity is a prerequisite for most forms of social valuation.
- 10.
- 11. **Michael Platt** (University of Pennsylvania)
 - **Contribution:** A versatile researcher studying the neural basis of social decision-making in both humans and primates. He has made key contributions to understanding social attention, vicarious reward, and the influence of hormones like oxytocin.
- 12.
- 13. **Fiery Cushman** (Harvard University)
 - **Contribution:** Blends moral psychology with reinforcement learning theory. He investigates how we assign credit and blame, and how model-free vs. model-based systems contribute to our moral judgments.
- 14.
- 15. **Patricia Lockwood** (University of Birmingham, UK)
 - **Contribution:** A rising leader in social neuroscience, focusing on the mechanisms of prosocial learning, empathy, and how these processes differ across the lifespan and in conditions like psychopathy and apathy.
- 16.

V. Attention, Memory & Other Cognitive Interfaces

Value doesn't exist in a vacuum. These researchers study how valuation is shaped by other fundamental cognitive systems.

1. **Ian Krajbich** (The Ohio State University)

- **Contribution:** A leader in using eye-tracking to inform models of choice. His work on the attentional drift-diffusion model (aDDM) shows that where we look, and for how long, directly influences the value signals that are computed in the brain.
- 2.
- 3. **Erie Boorman** (University of California, Davis)
 - **Contribution:** His research focuses on how we learn the hidden structure of decision problems, using this knowledge to guide choices. He investigates how the frontal cortex represents this "cognitive map" for valuation.
- 4.
- 5. **Dharshan Kumaran** (DeepMind / University College London)
 - **Contribution:** Works at the intersection of memory, learning, and decision-making. His research highlights the critical role of the hippocampus in organizing knowledge (e.g., social networks, conceptual structures) that supports flexible, model-based valuation.
- 6.
- 7. **Leor Hackel** (University of Southern California)
 - **Contribution:** An influential rising star who studies social learning. His work investigates the computational and neural mechanisms by which we learn about others' preferences and traits, and how we use that information to guide our own decisions.
- 8.
- 9. **Philip Corlett** (Yale University)
 - **Contribution:** A leader in computational psychiatry, exploring how aberrant "prediction errors" contribute to the formation and maintenance of delusions and other beliefs in psychosis. This provides a clinical window into how values and beliefs are encoded and updated.
- 1. **Stephen Fleming** (University College London)
 - **Contribution:** The world's leading researcher on the neuroscience and psychology of metacognition. He has developed influential computational models and experimental paradigms to show how the prefrontal cortex, particularly the anterior PFC, computes and represents confidence in our decisions, which is critical for learning and adapting behavior.
- 2. **Hakwan Lau** (RIKEN Center for Brain Science, Japan)
 - **Contribution:** A major figure in the neuroscience of consciousness and metacognition. His work uses a combination of psychophysics, fMRI, and causal

methods to dissect the neural circuits that support subjective confidence and awareness, distinguishing them from the circuits that perform the primary task.

3.

4. **Adam Kepecs** (Washington University in St. Louis)

- **Contribution:** A pioneer in studying the neural basis of confidence in rodents. His lab has identified specific neurons (e.g., in the orbitofrontal cortex) that signal decision confidence and has shown how these signals are used to guide behavior and learning, providing a cellular mechanism for metacognition.

5.

VII. Habit Formation & Instrumental Learning Circuits

Value encoding is not static; it drives the formation of habits. Understanding the transition from goal-directed action to habitual control is key.

1. **Ann Graybiel** (MIT)

- **Contribution:** A legendary neuroscientist whose work is foundational to our understanding of the basal ganglia. She discovered the "striosome" and "matrix" compartments of the striatum and has shown how these circuits are critical for learning, habit formation, and decision-making, especially in the context of cost-benefit analysis (e.g., approach-avoidance conflicts).

2.

3. **Bernard Balleine** (University of New South Wales, Australia)

- **Contribution:** A leading behavioral neuroscientist who has developed many of the core behavioral paradigms used to distinguish between goal-directed actions and habits. His work in rodents has been essential for mapping the specific corticostriatal circuits that mediate these two forms of value-based control.

4.

VIII. Affective Neuroscience & the Role of Emotion

Emotions are powerful valuation systems. These researchers study how core affective states shape choice.

1. **Elizabeth Phelps** (Harvard University)

- **Contribution:** A leader in studying the interaction of emotion and cognition. Her work has been pivotal in understanding how threat and fear (mediated by the amygdala)

influence learning, memory, and decision-making, and how these processes can be regulated by the prefrontal cortex.

2.

3. **Joseph LeDoux** (New York University)

- **Contribution:** A foundational figure in the neuroscience of fear. While his focus is on survival circuits, his work provides the essential framework for how the brain processes innate threats and learns new ones, which is a fundamental form of negative valuation.

4.

5. **Luiz Pessoa** (University of Maryland)

- **Contribution:** A prominent theorist who argues against a modular view of the brain. He advocates for the "dual competition" model, proposing that emotion, cognition, and motivation are deeply integrated and constantly compete for processing resources throughout the brain, challenging simple models of a central "value" hub.

6.

IX. Sensory Systems, Perception & Information

Value is not only attached to rewards but also to information itself. This research explores how value changes what we perceive and what we seek to know.

1. **Jacqueline Gottlieb** (Columbia University)

- **Contribution:** A leader in the neurophysiology of attention and curiosity. Her primate research has shown that the same brain regions that encode the value of rewards (like the LIP) also encode the value of *information*, providing a neural basis for curiosity and information-seeking behavior.

2.

3. **Shinsuke Shimojo** (Caltech)

- **Contribution:** A pioneer in studying the interface between perception, attention, and preference. His lab's elegant psychophysics experiments show that our subjective preferences and the values we assign to options can retroactively alter our conscious perception of them.

4.

5. **Kenji Doya** (OIST, Japan)

- **Contribution:** A major figure in computational neuroscience and robotics. His work has been crucial in conceptualizing the distinct roles of different basal ganglia loops and neuromodulators (dopamine, serotonin, norepinephrine) in reinforcement learning, covering prediction, timing, and overall goal management.

6.

X. Key Theorists & Modelers of Decision Dynamics

These researchers develop the high-level mathematical frameworks that are used to test theories of value accumulation and choice.

1. **Marius Usher** (Tel Aviv University, Israel)

- **Contribution:** A leading theorist in mathematical psychology and decision-making. He has developed sophisticated evidence-accumulation models (like the leaky competing accumulator model) that account for the dynamics of preference formation, attention, and context effects in choice.

2.

3. **Laurence T. Maloney** (New York University)

- **Contribution:** An expert in computational models of perception and motor control under uncertainty. He applies principles from economics and decision theory to understand how the brain handles risk and ambiguity in sensory and motor tasks, treating movement itself as a value-based decision.

4.

5. **Angela J. Yu** (UC San Diego)

- **Contribution:** A prominent computational theorist whose work focuses on normative Bayesian models of cognition. She has done influential work on the roles of acetylcholine and norepinephrine in signaling expected and unexpected uncertainty, which critically modulates how values are learned and relied upon.

6.

XI. Causality, Neuromodulation, and Whole-Brain Dynamics

Moving beyond correlation to understand circuit function and the role of key chemical modulators.

1. **Karl Deisseroth** (Stanford University)

- **Contribution:** A Nobel Prize-winning inventor of optogenetics and CLARITY. While not exclusively a value researcher, his techniques have revolutionized the field, allowing for the precise causal manipulation and mapping of the neural circuits (e.g., dopamine pathways) that are hypothesized to encode value.

2.

3. **Patricia H. Janak** (Johns Hopkins University)

- **Contribution:** A leading researcher on the cellular and circuit mechanisms of associative learning and addiction. Her work using electrophysiology and optogenetics in rodents precisely dissects how dopamine and other signals in the amygdala and striatum contribute to learning the value of reward-predicting cues.
- 4.
- 5. **Lior Appelbaum** (Bar-Ilan University, Israel)
 - **Contribution:** A pioneer in using the larval zebrafish model to study neuroscience. Due to the fish's transparency, he can perform whole-brain imaging at single-cell resolution during decision-making (like sleep/wake choice), providing an unprecedented system-level view of value-based state selection.
- 6.

China

China has invested heavily in neuroscience, and it has become a powerhouse in the study of the brain basis of value, social choice, and mental health.

1. **Xiaolin Zhou** (East China Normal University / Peking University)
 - **Contribution:** A leading figure in Chinese social and decision neuroscience. His lab has produced a large body of influential work on the neural underpinnings of fairness, empathy, social conformity, and altruism. He is particularly known for using fMRI and pharmacology (e.g., oxytocin) to understand how social norms and emotions modulate value signals.
- 2.
3. **Shihui Han** (Peking University)
 - **Contribution:** A global leader in cultural and social neuroscience. While not a traditional neuroeconomist, his work is fundamental to value encoding because he investigates how culture shapes the neural representation of the *self*. This self-representation is the anchor for subjective value. His research shows how cultural priorities (individualism vs. collectivism) alter activity in the vmPFC and other valuation regions.
- 4.
5. **Jian Li** (Peking University)
 - **Contribution:** A central figure in human neuroeconomics in China. His research focuses on the core mechanisms of value representation and choice, particularly the role of the orbitofrontal cortex (OFC) and vmPFC in encoding expected value, risk, and ambiguity. His work often involves elegant fMRI designs that test formal economic models.

6.

7. **Lusha Zhu** (Peking University)

- **Contribution:** A prominent researcher trained in the direct lineage of Paul Glimcher. She combines human fMRI, primate electrophysiology, and computational modeling to tackle foundational questions in neuroeconomics. Her work investigates how the brain represents and compares values under uncertainty and in social contexts, bridging the gap between human and animal models.

8.

9. **Shu Li** (Peking University / Institute of Psychology, Chinese Academy of Sciences)

- **Contribution:** A leader in computational psychiatry in China. He applies reinforcement learning (RL) and Bayesian models to understand decision-making deficits in clinical populations, especially individuals with substance use disorders. His work is critical for understanding how addiction hijacks the brain's value encoding systems.

10.

11. **Yi Rao** (Peking University)

- **Contribution:** A major figure in Chinese neuroscience, his work on the molecular and circuit basis of social behavior is highly relevant to value. He has done influential research on the role of serotonin in aggression and social dominance hierarchies, which can be understood as the valuation of social actions and status.

12.

Japan

Japan has a long and storied history in neuroscience and robotics, which has created a unique environment for computational and systems-level approaches to value.

1. **Kenji Doya** (Okinawa Institute of Science and Technology - OIST)

- **Contribution:** A world-leading computational neuroscientist. His work provides the theoretical foundation for how different brain systems and neuromodulators handle different aspects of reinforcement learning. His framework conceptualizing the distinct roles of dopamine (reward prediction), serotonin (regulating patience/impulsivity), and norepinephrine (managing uncertainty) is foundational to the field.

2.

3. **Masamichi Sakagami** (Tamagawa University)

- **Contribution:** A key researcher in primate neurophysiology. His work focuses on how the prefrontal cortex, particularly the orbitofrontal cortex (OFC) and lateral PFC,

flexibly encode and update value signals based on changing rules and contexts. He is a leader in understanding the neural basis of cognitive flexibility in economic choice.

4.

5. **Hakwan Lau** (RIKEN Center for Brain Science)

- **Contribution:** A global leader in the neuroscience of metacognition and consciousness. His work is crucial for understanding how the brain represents *confidence* in a value judgment. He investigates the prefrontal circuits that allow us to know what we know, which is a critical, second-order value signal.

6.

7. **Shinsuke Shimojo** (University of Tokyo, WPI-IRCIN)

- **Contribution:** A pioneer in studying the powerful, bidirectional link between perception, preference, and value. His elegant psychophysics experiments demonstrate that our subjective preferences can physically alter what we perceive, suggesting that value is not just a post-hoc tag but a fundamental part of sensory processing.

8.

9. **Keise Izuma** (Kochi University of Technology)

- **Contribution:** An influential social neuroscientist known for his work on social influence and reputation. His fMRI studies investigate how the brain's valuation circuitry (e.g., striatum) responds to social rewards like a good reputation and how brain activity changes when we conform to the opinions of others.

10.

South Korea

South Korea has developed a strong focus on high-tech, translational neuroscience, with a particular strength in computational psychiatry and linking brain signals to real-world behavior.

1. **Hackjin Kim** (Korea University)

- **Contribution:** A prominent figure in affective and decision neuroscience. His research explores how negative emotions like anxiety and threat fundamentally alter value-based decision-making. He is known for his work detailing the roles of the amygdala, insula, and prefrontal cortex in processing risk, uncertainty, and aversive outcomes.

2.

3. **Woo-Young Ahn** (Seoul National University)

- **Contribution:** A leading voice in computational psychiatry. He uses sophisticated machine learning and hierarchical Bayesian modeling to create "computational phenotypes" of mental illness. His work is at the forefront of using decision-making data to predict substance use relapse and develop personalized treatments, directly linking value-encoding models to clinical outcomes.
- 4.
- 5. **Sang-Hun Lee (KAIST)**
 - **Contribution:** An expert on the interplay between vision, attention, and reward. His research demonstrates that the brain's valuation systems don't just wait for information; they actively modulate how we process sensory input. He has shown how reward expectations can change neural activity in the earliest stages of the visual cortex, sharpening perception for valuable stimuli.

Singapore

As a global research hub, Singapore has attracted top talent and fostered a highly interdisciplinary environment for neuroscience.

1. **O'Dhaniel Mullette-Gillman (National University of Singapore - NUS)**
 - **Contribution:** A leader in the field of developmental neuroeconomics. His research investigates how value-based decision-making, particularly concerning risk and reward, changes across the lifespan from childhood through adolescence to adulthood. His work helps explain why adolescents are often more prone to risky behavior from a neuro-maturational perspective.
- 2.
3. **Juan Helen Zhou (National University of Singapore - NUS)**
 - **Contribution:** An expert in brain imaging and network neuroscience. While her focus is often on aging and neurodegenerative disease, her work provides a critical systems-level perspective. She studies how large-scale brain networks (like the default mode network, which overlaps with valuation areas) support cognition and how their breakdown impairs functions like decision-making.
- 4.

Companies we'd like to partner with for data:

- Brain scanning companies

- fmir

Guides

Agent Foundations Theory:

Essential skills/knowledge - **Measure theory/real analysis**

How this aims to solve alignment:

Current State of Research: Understanding Trust by Abram Demski - the Tiling Problem, the Learning Theoretic Agenda

Bottlenecks suggested by Senior Researchers:

For the Learning Theoretic Agenda

Bottlenecks found by AI Plans researchers in preparation for this: Everyone's siloed - working in their own area

Vanessa Kosoy's thing is quite different from Abram Demski's, which is different from Sahils, etc

Agent Foundations Applied

Essential skills/knowledge

How this aims to solve alignment:

Current State of Research: Most Agent Foundations methods

Bottlenecks suggested by Senior Researchers:

Neuroscience Based AI Alignment

Essential skills/knowledge

How this aims to solve alignment:

Existing Literature:

Bottlenecks suggested by Senior Researchers: fmri data is noisy

Improved Preference Optimization

Essential skills/knowledge: DL, ML

How this aims to solve alignment: do things like RLAI, DPO, PPO, etc, but more robustly, in ways that scale more, have more evidence of scaling - maybe using interp based pref optimization,

Existing Literature:

Bottlenecks suggested by Senior Researchers:

General Bottlenecks that everything has:

Not precise, unambiguous, robust to scale, alignment evals

Knowing exactly what to measure in alignment evals

How to get GPUs:

https://docs.google.com/document/d/1dqsUPoJRjX_dRe-ByDcDLpasxRp-XmWE162LiDPIAo/edit?tab=t.0#heading=h.ni2mkpurj7yt

Neuroscience

Computational RL would be useful for analyzing fMRI data

Agent Foundations

Tests

Theoretical Agent Foundations

What is Decision Theory?

- A. How to make good decisions
- B. a quack science
- C. a nice book
- D. A branch of maths, about decisions

Improved Preference Optimization

Need to know: what a reward model is, how preference optimization works, what goes into making the architecture of this, what some errors/mistakes are to look out for in this, test for specific details and stuff here - just making sure they know what they're talking about when it comes to making a new method of preference optimization

Event Schedule

Explainer for: <https://lu.ma/8ehdokx7>

The hard part of alignment

How could we make a system of dramatically superhuman competence that's pointed at good things, and whose future iterations all remain pointed at good things?

How do we make a system that is more competent than all human beings in all tasks and make sure it completes tasks in ways that are universally agreed to be beneficial to human beings?

How do we make sure the system's future versions all act in ways that are only beneficial to human beings?

Week 1

lit review plus idea submission - main goal here is to really understand the problem, what work has already been done and what they can do to make progress on the problem.

also, giving feedback to fellow participants.

current idea: this will be done with the Brainstorming channel on the discord, where they'll get feedback from me and other research assistants.

what they should do after getting the pass from stage 1:

- make introduction in the #introductions channel
- if they have an idea, then submit that idea in the Brainstorming channel
 - we'll also submit some idea in the Brainstorming channel ourselves

First week will be literature review and exploring data on what's been done, what's failed, what they can improve on, what resources they'll need, etc - and if they have a relevant mentor, then coordinating with their mentor on this

by the end of the first week, they should have a thorough understanding of everything that's been attempted that's been similar or linked to their idea before - they should know why this architecture hasn't been tried before, they should know what might make it fail, what would be signs of it failing, what they need to draw on and get ideas from, in order to make it something that will get values into the model - they should have some idea of the evidence they'll have for the values being in the model

if their track is agent foundations, the evidence might be something like a formal proof in a situation with some assumptions, or some closed case/environment - and they've got a hypothesis/idea for how to extend that into an architecture

if their track is brain based ai safety, then it might be some evidence/group of studies, that seem to find a particular part/parts of the brain that's responsible for encoding values - and perhaps some idea of how this part of the brain forms/develops - and they have some alignment evals, interpretability methods, etc, of checking their methods will work

if their track is preference optimization then it might be evidence of how it's different from other methods and some mathematical theory, evidence of what's worked better/worse in previous research papers, alignment evals, interpretability methods, robustness tests, etc

If track is original/other, should also include evidence for why the architecture will be able to scale in capabilities as well, alongside the alignment evals, interpretability, robustness test, etc

Week 2 - 3

Coming up with experiments, based on what they've learnt, discussed with mentors, etc, and executing - they'll have access to compute in the form of TPU credits, advice on how to get additional compute and advice from AI researchers on the AI Plans team

Then, executing their experiments, iterating on them, running them on larger scale models, or if doing a pre training method rather than a post training method, running with more parameters.

They should also be frequently running alignment evals on their models, seeing how scores change, checking for potential reward hacking of the evals (this may be tricky and uncertain).

For the agent foundations teams, I'll have more specific advice tomorrow (12th June) (have a call with a senior agent foundations researcher tomorrow).

(after the call)

- 2 sub tracks
 - Theoretical focus
 - Category Theory: String Machines, String Diagrams "diagrammatic notation for constructing new infrakernels out of given infrakernels."
 -
 -

Week 4

Thoroughly checking and trying to critique and red team the methods they used to verify the alignment of their models and if their alignment methods worked. Also, testing this on larger scale as well. AI Plans team will provide thorough advice on this, since we have experience with reward hacking evals.

Week 5

Write up their paper, methodology, limitations, etc. And also prepare their Poster for the Presentation Evening

Final Day

Presentation Evening (online) - this will likely be held on Gather Town, though I still need to talk with them to confirm it and a teammate suggested a different site may also work. Teams will have little online places where they can present their work, people will ask questions, etc. Senior researchers will be looking at posters and vote on a winner.

Job Fair (online) - various companies, such as DeepMind (not certain, need to confirm - but I know a few people working there and think they'd be interested), Prism Evals, HUD Evals, etc will have booths, talk about roles they're looking for, etc

Mentors

Cole Wyeth

Agent Foundations Theory

Working on AIXI with Marcus Hutter

Research Managers

This will be general people with alignment research experience that people taking part in the program can call upon, have calls with, ask questions, etc

Péter Trócsányi

Potential Mentors:

Abram Demski

Tiling Problem

Agent

Speakers

Speakers Secured:

Abram Demski - probably, he's said yes - could end up being busy though
Quentin (prism evals co founder) - probably

Speakers Desired

Nate Soares
Yudkowsky (maybe, could be bad)
Joshua Greene (cogsci researcher)
Rebecca Weber (cogsci)
Nathan Lambert
Other Preference Optimization researchers

Speakers Declined (they said no)

Vanessa Kosoy - said yes before

people to find

mathematics phd students at top universities

- also postdocs and professors
- maybe unusually brilliant masters student - but this is hard to gauge

people who specialize in:

- logic,
- computability
- probability
- computational complexity
- could be really theoretical comp sci phds as well

get senior ex miri people to formulate mathematical questions that are blocking their agendas

- converse laverie conjecture
(<https://www.lesswrong.com/posts/5bd75cc58225bf06703753b9/the-ubiquitous-converse-lawvere-problem>)
-
- bring one or two agent foundations researcher, to provide context, ask questions, etc

Luma Page

Alignment Moonshot: 30-Day Intensive Research Program

A 30-day program to develop falsifiable methods for embedding values into AI models that generalize and scale.

Most alignment research focuses on subproblems. This program tackles the core challenge directly: getting values into models with strong empirical evidence that the methods work.

Program Details

Duration: 30 days full-time **Format:** Teams of 3-5 researchers **Start:** Kickoff call with Kabir Kumar **End:** Poster Evening followed by Careers Fair

Research Tracks

1. Brain-Based AI Safety

Develop architectures based on how morals/values are encoded in the human brain.

2. Improved Preference Optimization Methods

Create non-shallow methods of scalable oversight that demonstrably embed values deeply into models. Must show generalization beyond training distribution through interpretability-based evaluations.

3. Agent Foundations

Either solve mathematical problems in agent foundations or implement existing theoretical work like Infrabayesianism.

4. Original/Other Methods

Novel approaches to the core alignment problem that don't fit other tracks.

Program Structure

Week 1: Form teams, define specific approaches, design falsifiable experiments **Week 2:** Build implementations, run initial experiments, iterate daily **Week 3:** Test for generalization, scale to larger models, peer review **Week 4:** Finalize work, prepare posters, document methods

What We Provide

- GPU compute for experiments
- Mentorship from senior alignment researchers
- Collaboration infrastructure
- Stipends for full-time participants
- Presentation opportunity to AI lab representatives

Application Process

Stage 1 requirements:

1. CV
2. Confidence level (1-10) you can commit for 30 days
3. Brief explanation of why you can commit
4. (Optional) Additional relevant work/ideas/links

Final Events

Poster Evening

Teams present their work in Gather Town. Senior alignment researchers judge projects. Conference-style format where attendees can visit your virtual booth.

Careers Fair

Representatives from DeepMind, Anthropic, Redwood Research, Apart Research, MIRI, Conjecture, CAIS, FAR AI, Ought, and others will discuss opportunities at their organizations.

Expected Outcomes

- Working implementations of value alignment techniques
- Empirical evidence of generalization and scaling
- Falsifiable predictions with test results
- Open-source contributions
- Direct connections to alignment organizations

Questions? Contact kabir@ai-plans.com

Agent Foundations Notes

Explainer for: <https://lu.ma/8ehdokx7>

The hard part of alignment

How could we make a system of dramatically superhuman competence that's pointed at good things, and whose future iterations all remain pointed at good things?

What we're currently doing:

The current training methods involve using data to direct an optimizer to find a minima in the probability distribution of possible weights of a neural network. To find a minima that leads The current training method is finding the set of weights whose outputs have the lowest difference between the output that we're looking for and the output that we get, without getting too close and overfitting.

loss when compared to a dataset. Today's loss functions typically guide a model's behavior towards a dataset. Reward functions also rate the data itself and almost always involve collecting new data. Both rely on the model empirically generalizing well.

Why it's not going to scale:

This works temporarily, but as the world's distribution changes, even if the model is successfully learning what the rating functions have to tell it, it eventually fails to do what the creator wanted.

Giving feedback on which outcomes are 'good' or 'bad' requires a judge which knows what 'good' and 'bad' are, so a raw environment isn't able to do this, unlike giving feedback on if a program will run or not. Similarly, this may be why basic RL leads to increase in performance for math and coding for LLMs, but when scaled, a decrease in performance of writing poetry:

<https://aidanmclaughlin.notion.site/reasoners-problem>

What process encodes generally doing good things across wide contexts. What we hope to find

What we need to solve for a scalable solution:

At some point we are going to have extremely competent AI systems which can solve problems including, but not limited to, "how to make a more capable AI," "how to be extremely persuasive," etc.

When this happens, we want to ensure that this AI and all future AIs that it might make will be empowering and helping humanity, rather than harming or disempowering humanity. To do this, we need to be certain of the preferences of the AI and how it will act on those preferences.

How do we *be certain* that the AI will have the preferences needed for good things to happen and not have any potential for harm?

To do this, we need to solve some very challenging problems that involve not just the standard programming, linear algebra & calculus problems, but also decision theory, game theory and other less well known topics.

Some terms used in this document may be new or ambiguous. Please see the Glossary section at the bottom for meanings of terms such as 'formally specify in math' - or feel free to add a comment. All comments and questions are welcome - don't worry about the questions being too 'basic' - happy to help.

The Tiling Problem:

How to make sure that AIs made by a 'good' AI will continue to be as 'good' and that AIs made by those AIs, will be 'good', then AIs made by those, etc

How do we ensure that AI systems created by aligned AIs remain aligned, and so on through multiple generations of systems?

Skills involved: Strong math background, ability to read [dense mathematical papers](#), learn new branches of mathematics, decision theory, game theory

Resources:

[updateless decision theory 1.1](#) tiles

Problems

- requires omniscience
-

Understanding Trust by Abram Demski

<https://drive.google.com/file/d/1wsAND3AqnGsPCCWor1OKcBX-pnMCSuoP/view>

How to get values into the AI

A major part of the alignment problem is making sure that the exact goal we choose is the exact goal that the AI will pursue.

Making sure that the goal is followed by all future AIs, made by AIs, is difficult.

Difficulties Involved:

[Alignment is in large part about making cognition aimable over significant durations at all.](#)

If you're interested in working on this, please let me know, I'd want to work closely with you - there are ways to do this successfully, and there are ways to get halfway there and end up worse than how you started.

Alignment Target Ideation

What potentially mathematizable / formally specifiable target could produce good outcomes even if optimized for extremely strongly? Must not fail to [Goodhart's Curse](#).

Which goals can we formally specify with math, that will not lead to bad outcomes if over-optimized for. In other words, consider today's training targets. Today's loss functions typically guide a model's behavior towards a dataset. Reward functions also rate the data itself and almost always involve collecting new data. both rely on the model empirically generalizing well. This works temporarily, but as the world's distribution changes, even if the model is successfully learning what the rating functions have to tell it, it eventually fails to do what the author wanted. A well specified target is one that will keep producing ratings relative to something reasonable no matter how strong the model gets, even if the author can't edit it anymore.

E.g. the goal of 'reducing human suffering' could be optimized by reducing the human population to 0; If the model is initialized

Attempts at solving the problem:

QACI by Orthogonal

Problems with the resources:

- Strange grammar
- Either insufficient math or a lot of math that is the wrong math specified inconsistently

The Value to get out of them:

- if you can see through the way the wrong math is being used inconsistently and the unnecessary new terms, you can get an idea why they were trying to do it this way
- You can

Predca by Vanessa .

Physicalist Superimitation by Vanessa Kosoy

https://www.lesswrong.com/posts/ZwshvqiqCvXPszEct/the-learning-theoretic-agenda-status-2023#Physicalist_Superimitation

Skills involved: Creativity, clear headed and precise philosophy,

Problems with the resources:

Difficulties: In addition to the problem being extremely math dense, the resources for learning about it are extremely math dense

Sidestepping the need to solve morality?

This may be a sub problem of Alignment Target Ideation. We may need to find out how to dodge this, or even just solve this thousands-year-old-problem this decade, as opposed to just hoping it isn't needed. It's probably not the most tractable, but we didn't feel right not mentioning it.

Potential Ideas/attempts on dodging the problem:

Problems with Understanding Agents

How is it possible for agents to behave intelligently, even within computational limits. How do you make a capable agent without a neural network - what is the mathematics that allows neural networks/a capable agent to exist

Realizability

Should not 'just catch fire' if it sees something it previously believed to be impossible

Embeddedness

Vingean Uncertainty

Logical Uncertainty

Ontology Mismatch

Having things classified/prioritized in the wrong order of importance

Corrigibility

How willing the AGI is to let us change it's goals/preferences.

Automating research for any of the above problems with AIs

A team of AIs might help solve the above problems and make parts of them easier to solve. However, having it be a team of AIs could introduce new problems and ambiguity. It could also make previous problems more difficult.

In addition to the previous difficulties of the problems, what new difficulties might arise from it being a team of AIs rather than humans?

When might there be more

Knowing the shape of the problem

Knowing what success means

What kind of ideas might be useful for that

[Agent Foundations for Superintelligence-Robust Alignment](#) has a bunch of suggested readings if you'd like to explore.

[Arbital](#) has a lot of relevant content.

Prizes

1. Regular: for making a very good idea/plan
2. Regular: for making very good critiques, finding vulnerabilities very well, in plan/idea
3. Super Combo: Both great idea **and** vulnerabilities of idea

Useful Resources

An attempt at a robust pointer to human values: Universal Alignment Test

<https://docs.google.com/document/d/1CMTS36MCbykYirTmC9PdI2RBqLLPmrFU1sDcBNMvDCk/edit?tab=t.0>

Glossary

'formally specify in math'

we need a way to specify this that is robust to the fact that we haven't finished science on physics - we don't have a working theory of everything and don't expect to for a while, and that needs to not matter when the AI gets capable enough to do fundamental physics research on its own. (Though that seems like it might take a while, compared to other concerns from drastically superhuman models doing science.)

merely math that

Message from the Alignment Evals Hackathon:

Hello __,

I'm hosting an AI Alignment Evals hackathon on the 25th of January.
Will include learning sessions on how to make and use safety benchmarks for Blue Teams, how to finetune models to deceptively fit them for Red Teams.
Would you consider sharing this with folks at the ____?

the message could be something like:

...

AI-Plans is hosting an AI Alignment Evals Hackathon on January 25th
Join a Blue Team to make the next benchmark for AI Alignment or a Red Team to make the model that breaks it

Blue teams will learn:

How to make an alignment benchmark

How to get scores from benchmarks

How to use Inspect, an evals framework by the UK AISI

What makes a benchmark robust to targeting

Red Teams will learn:

How to finetune a models to pass alignment benchmarks, without actually having the desired values

How to make a small 'aligner' model that makes the model output look aligned, without the model itself being aligned.

Featuring talks from:

Monika Jotautaitė, Technical Project Manager at the AI Safety Engineering Taskforce.

James Hindmarch, co-author of the ARENA evals course.

Zainab Ali Majid, co-author of Rethinking CyberSecEval: An LLM-Aided Approach to Evaluation Critique.

Sign up here before January 20th: <https://lu.ma/xjxqcy>

...

Best,

Kabir Kumar,

Founder, AI-Plans

Preference Optimization

Scheduling/Organizing

Application Process

the application process:

Stage 1: CV, % likelihood of being able to do 10 hours a week for 5 weeks from August 9th, which tracks you're interested in. Optional - anything else you think is relevant which is not in your CV

Stage 2: basic knowledge check for the tracks you're interested in - 15 timed multiple choice questions - e.g. for agent foundations, will be basic theoretical computer science, probability theory, decision theory, etc, for neuroscience track, basic neuroscience, fMRI, modelling, etc. Also, a Yes/No for if you'd be interested in being a Team Leader

Stage 3 - you'll get invited to a discord, along with all the other applicants who've made it to this stage - form a team, agree on an idea and submit it - you'll get access to concise resources on the current methods and bottlenecks in all the tracks, which we've made by interviewing lots of senior researchers.

For Stage 1, we guarantee personalized feedback to the first 300 applicants

For Stage 3, we guarantee feedback to the first 100 teams to submit an idea

Gather

**MAKE SURE THE STAGE CAN BE KEPT
EXCLUSIVE - SO THAT PEOPLE WHO HAVEN'T
PAID, CANT GO ON STAGE - WAS A PROBLEM IN
SPAR!! WILL BE \$2000 WASTED FOR
COMPANIES IF NOT!!!**

Message From Sasha:

Hi!

i'm organising an ai alignment research conference

Kabir Kumar7:42 PM

going to have a poster day

do you know what the limit is in the number of people that can be in a space?

if email is better, btw, my email is kabir@ai-plans.com

May 27, 2025

Sasha (EAGather steward)2:36 AM

Hi Kabir

Sasha (EAGather steward)2:50 AM

The Gather currently has a paid capacity of 60 people, so if your event will comfortably under that you can use it for free. If you think it might be more than that, the cost per extra user is something like \$2.10 per user per day capped at \$4.9 per user per day.* If you do want to increase it I'll put you in contact with Anima International since they're paying for the current capacity

* I suggest if you do want to raise the cap offering to pay an extra \$50-100 toward sharing the base cost for the period of the initial 60, since a) you're co-benefiting with Anima from the initial 60 users they've already paid for and b) they'll have to deal with a bit of extra admin to increase the capacity for you**

** I don't have any affiliation with Anima, and as far as I know they don't formally ask for any extra money in these cases - it just feels like good incentives :)

Sasha (EAGather steward)3:12 AM

btw, I've also been in touch with someone called Eitan re something similar. Are you guys from the same org?

June 18, 2025

Kabir Kumar1:34 PM

Hi, that would be great!! thank you!

Kabir Kumar1:35 PM

Sorry for the late reply!

We're not, but we actually met at the EAG Afterparty at LISA!

Kabir Kumar1:36 PM

Currently I'm thinking the number of people at the Poster Day might be something like 500 to 1000

Sasha (EAGather steward)4:30 PM

I think Gather has a limit of 500 even paid, though I'd have to check
you could potentially have two linked spaces if that's not enough

June 21, 2025

Kabir Kumar6:52 PM

hmm, i'm going to need to email Gather, in that case - thank you!

June 22, 2025

Sasha (EAGather steward)6:47 AM

Do you want me to put you in touch with Ania at Anima?

July 05, 2025

Kabir Kumar3:49 PM

Hi! Sorry for the late reply! Yes please!

Sasha (EAGather steward)3:49 PM

what's your email address?

Kabir Kumar4:03 PM

kabir@ai-plans.com

Emails to Participants (Drafts)

Stage 1 Accept

Thank you for applying to the Moonshot Alignment Program! You've been accepted! The feedback on your application was:

"

"

The [next steps](#) are:

Join the discord:

Make an introduction in the #introduction

I'm sorry it's taken so long to get feedback to you. Going through 298 applications by myself and giving feedback to each of them, while running also AI Plans took a lot more time than I thought - delegating it was trickier than anticipated, because its a really important step that requires really understanding the research and what's needed very deeply. And skills like judging when someone is actually a good candidate but bad at communicating themselves, or someone very impressive looking but either not relevant or better at yapping than actual work/research.

We're going to make this much smoother and faster for the next research program! I'm glad to say that We'll be guaranteeing feedback to just 50 applicants and starting early.

Best,
Kabir Kumar,
Founder, AI Plans

Hi (name),

Thank you for applying to the Moonshot Alignment Program! You've been accepted! After reviewing your application, I was impressed by (feedback) in your application.

OR

Hi (name),

You've been accepted to the Moonshot Alignment Program!

Your Application Feedback was:

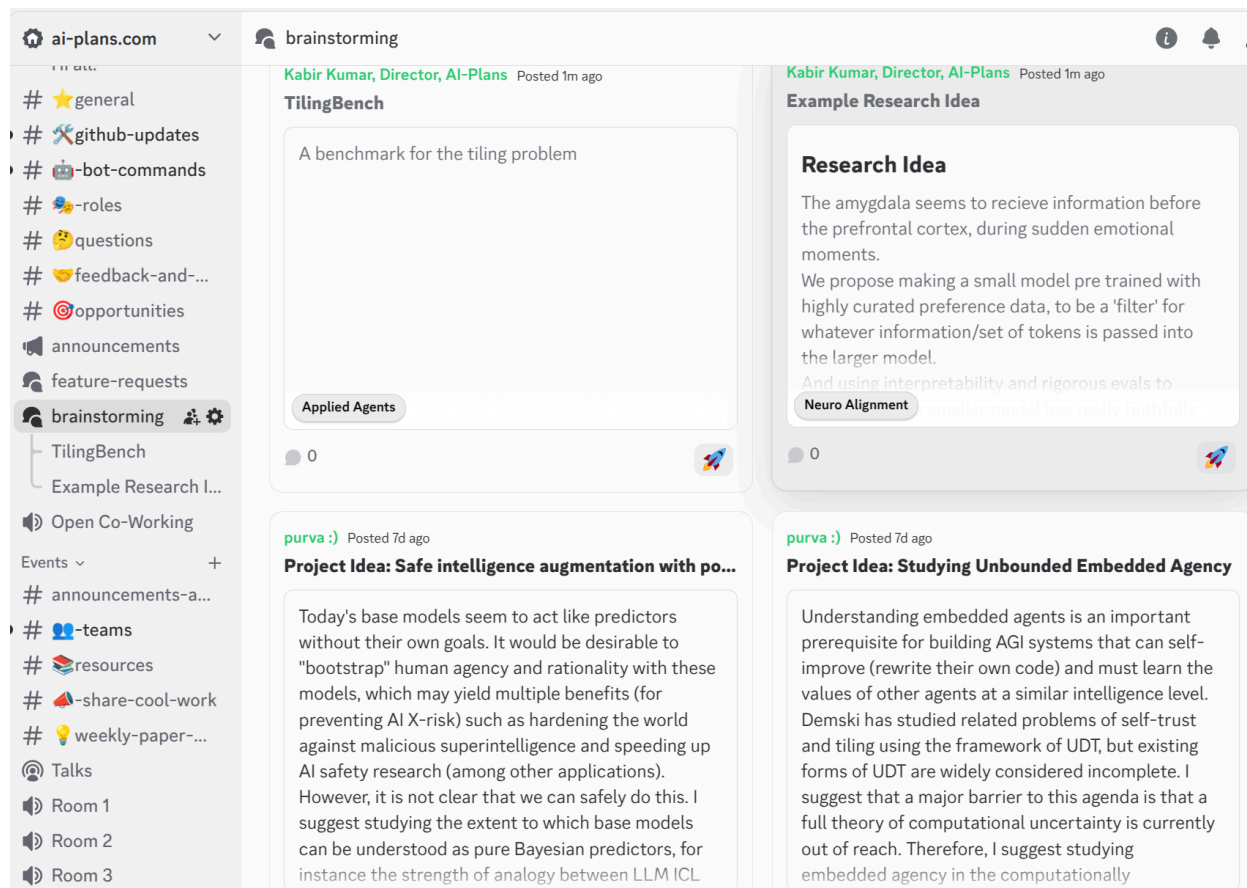
(insert feedback)

Next Steps

All the details for getting started are [here](#)

Due to the delay in getting started we will be skipping the Scaling part of the schedule, instead the schedule will be:

Week 1, August 13th to 19th: Literature review and Experiment Planning



Resources: The Research Guides, the Broad Vulnerabilities list, Research Assistants

Week 2, August 19th to 26th: Experiment Implementation and Testing

Week 3, August 26th to September 1st: Writeup and Planning of Extending Research

September 2nd: Presentation Day and Job Fair

Week

A note on the process

I'm really sorry for taking so long to get back to you. Reviewing 298 applications by myself while also managing [AI-Plans](#) took much longer than expected. Making the Research Guides has also been harder than I thought it would be. We've interviewed lots of senior alignment researchers to learn about their research process, to make these guides, but it also involved a lot of iteration,

learning the research ourselves and figuring out how to state the problems more concisely and with less jargon, which was hard. Any feedback on the guides, including critical, is appreciated.

We're already planning improvements for our next program and we won't be starting from scratch next time, so we'll be able to do better

Best,
Kabir Kumar,
Founder, AI Plans

"

The [next steps](#) are:

Join the discord:

Make an introduction in the #introduction

I'm sorry it's taken so long to get feedback to you. Going through 298 applications by myself and giving feedback to each of them, while running also AI Plans took a lot more time than I thought - delegating it was trickier than anticipated, because it's a really important step that requires really understanding the research and what's needed very deeply. And skills like judging when someone is actually a good candidate but bad at communicating themselves, or someone very impressive looking but either not relevant or better at yapping than actual work/research.

We're going to make this much smoother and faster for the next research program! We'll be guaranteeing feedback to just 50 applicants and starting early.

Best,
Kabir Kumar,
Founder, AI Plans

P.S. We've currently had \$5

Stage 1 Reject

Thank you for applying to the Moonshot Alignment Program! Unfortunately, you didn't get accepted from Stage 1. The feedback on your application was:

" "

You're very welcome to take part in the coming events which are open to everyone:

<https://lu.ma/ai-plans>

There is also the AI Law event, which is open to those with experience in Law and Policy, as well as for researchers.

I'm sorry it's taken so long to get feedback to you. Going through 298 applications by myself and giving feedback to each of them, while also running AI Plans took a lot more time than I thought - delegating it was trickier than anticipated, because it's a really important step that requires really understanding the research and what's needed very deeply. And skills like judging when someone is actually a good candidate but bad at communicating themselves, or someone very impressive looking but either not relevant or better at yapping than actual work/research.

We're going to make this much smoother and faster for the next research program! We'll be guaranteeing feedback to just 50 applicants and starting early.

Best,
Kabir Kumar,
Founder, AI Plans

Tab 35

Stage 1 Accept

Hi (name),

You've been accepted to the Moonshot Alignment Program!

Your Application Feedback was:

(insert feedback)

Next Steps

All the details for getting started are [here](#)

Due to the delay in getting started we will be skipping the Scaling part of the schedule, instead the schedule will be:

Week 1, August 13th to 19th: Literature review and Experiment Planning

Week 2, August 19th to 26th: Experiment Implementation and Testing

Week 3, August 26th to September 1st: Writeup and Planning of Extending Research

September 2nd: Presentation Day and Job Fair

A note on the process

I'm really sorry for taking so long to get back to you. Reviewing 298 applications by myself while also managing [AI-Plans](#) took much longer than expected. Making the Research Guides has also been harder than I thought it would be. We've interviewed lots of senior alignment researchers to learn about their research process, to make these guides, but it also involved a lot of iteration, learning the research ourselves and figuring out how to state the problems more concisely and with less jargon, which was hard. Any feedback on the guides, including critical, is appreciated.

We're already planning improvements for our next program and we won't be starting from scratch next time, so we'll be able to do better

Best,
Kabir Kumar,
Founder, AI Plans

"

The [next steps](#) are:

Join the discord:

Make an introduction in the #introduction

I'm sorry it's taken so long to get feedback to you. Going through 298 applications by myself and giving feedback to each of them, while running also AI Plans took a lot more time than I thought - delegating it was trickier than anticipated, because it's a really important step that requires really understanding the research and what's needed very deeply. And skills like judging when someone is actually a good candidate but bad at communicating themselves, or someone very impressive looking but either not relevant or better at yapping than actual work/research.

We're going to make this much smoother and faster for the next research program! We'll be guaranteeing feedback to just 50 applicants and starting early.

Best,
Kabir Kumar,
Founder, AI Plans

P.S. We've currently had \$5

Stage 1 Reject

Thank you for applying to the Moonshot Alignment Program! Unfortunately, you didn't get accepted from Stage 1. The feedback on your application was:

" "

You're very welcome to take part in the coming events which are open to everyone:

<https://lu.ma/ai-plans>

There is also the AI Law event, which is open to those with experience in Law and Policy, as well as for researchers.

I'm sorry it's taken so long to get feedback to you. Going through 298 applications by myself and giving feedback to each of them, while also running AI Plans took a lot more time than I thought - delegating it was trickier than anticipated, because it's a really important step that requires really understanding the research and what's needed very deeply. And skills like judging when someone is actually a good candidate but bad at communicating themselves, or someone very impressive looking but either not relevant or better at yapping than actual work/research.

We're going to make this much smoother and faster for the next research program! We'll be guaranteeing feedback to just 50 applicants and starting early.

Best,
Kabir Kumar,
Founder, AI Plans

Improvements for Next Time

Have the team building session in gathertown, have people wander around, have them commit to taking part in the launch session, so they form teams, talk to each other, etc.

They can use discord first if they want.

But the commitment is that by the end of the launch session they should have a team.

At 200+ people, lots of people saying "i want a teammate" , " i already have a teammate", "i dont know if i can be on your team", etc

What tracks they want to do can be a badge in discord and on gathertown

- in Gathertown, can be like Name (Neuroscience), Name (theory agent), etc

Have a self hosted Gathertown alternative, or funds for gathertown

Ask Senior Researchers, what their team roles/structures are, how their project management works, etc. Especially for the Theory researchers - how is work divided, who does what, how is this decided, how is work kept track of?

Ho