Ambitious Meta/Cybercognition

For the past year, I've been working on a new rationality training, focused on the particular domain of "solve confusing problems at the edge of your intellectual abilities)". The main mechanisms I'm focusing on, that seem distinct from previous projects, are:

- Inventing better feedback-loops *(faster, richer, more accurate, more relevant)*
- Emphasis of deliberate practice *(i.e. it's not surprising if you aren't permanently more rational after a weekend workshop, but it's surprising to me if 100 hours of solid practice at the edge of your ability doesn't help in obvious ways)*

More recently, I've also been working on cyborg tools and using LLMs to augment my cognition. I plan to scale some of those tools as public LW features, and some as custom powertools for individuals that seem worth differentially accelerating.

I currently see these as part of a cohesive plan. I think they will likely pay off in the nearterm. And, I hope, more ambitiously, to go for the throat on "olympic level training and AI exobrain tooling for x-risk researchers" (with a particular hope that this can enable a lot of actual alignment research, focused on the hard parts of the AI problem).

The most ambitious dream is that, by the time "golden hour" comes (i.e. the period where AI is as capable as it can be without being unsafe), there is a cohort of 100+ alignment researchers, focused on the most important parts of the problem, with cognitive tools to deal with it's difficulties, and fluency it outsourcing the parts of cognition that are safe and helpful to outsource to AI.

But for now, I'm structuring my plans so they pay off along the way, and hopefully resolve cruxes early about whether the more ambitious hopes will work.

I'm also mostly trying to structure them in a way that doesn't depend on the limited supply of mentorship (i.e. get people a bunch of obvious skills, with obvious feedbackloops, that hopefully Nate, Eliezer, Anna, Paul Christiano, Chris Olah, Evan Hubinger, and Buck would all look at and say "okay that seems like it'd at least reasonably help")

The main things I've done/am-doing so far

2.5 day Cognitive Bootcamps, so far charging $600, planning to charge $1000 for the next ones, and raising prices over time as I become more confident in them.

Monthly "Recursive Self Improvement Club", for workshop alumni, to continue solidifying their skills and explore new ones. Iterating on what sort of followup exercises meaningfully help people.

Iterating on Lightcone team, aiming in 12 months to 2x our end-to-end velocity (which should be pretty obvious if it works) and also hopefully 2+x our strategic aim, and ability to grapple with confusing problems (which will be less obvious, but I'll do my best improving feedback loops about that).

Iterating on Timaeus, a Singular Learning theory org (aiming to both help them increase their velocity, and also get them on board with "have at least 2 major backup plans that aren't Singular Learning Theory, and 2 backup yearlong plans that aren't their current specific agenda." Two Timaeus employees have done the workshop, and I'll be chatting with a higher-up soon.

Developing LLM-powered tools that improve LessWrong, or our internal team process.

My own experience applying this to myself

There's a failure mode where the people who self-select into "rationality teacher" are sort of mid-level and not really the most competent people around. I… notably am also one of those people. But, I'm basically setting my own standard here as:

- I plan to spend at least half my time actually solving hard problems at the edge of my ability, that force me to develop new and better techniques to succeed.
- Every 1-4 weeks, I should have:
    - Shipped a concrete thing publicly
    - Asked myself twice daily "Am I cutting towards the most important goal right now? How can I cut my hypothesis space in half as fast as possible?"

    - Each day, spend at least a few minutes (hopefully more) explicitly "deliberate practicing" something on the edge of my ability.
    - Allocated substantial attention towards:
        - "am I backchaining realistically from 'end the acute risk period?'"
        - "am I forward chaining tractably towards things that will pay off soon?"
        - "being wholesome". (i.e. the whole thing feels holistically good, not just minmaxed in brittle ways. And, I'm happy, and healthy)
    - There should be new classes of thoughts that take noticeably less time than they did a month ago.

My own optimism here routes largely through "Well, it seems like I'm just in fact way smarter than I was 14 years ago, and it is mostly through simple, straightforward habits that I learned in a piecemeal, scattered way. Surely a dedicate training regime version of it would be better?"

14 years ago, I couldn't think "strategically" (i.e. modeling goals/consequences in ways that stretched my working memory) for 5 minutes without getting a headache. Recently I spent basically two weeks persistently doing so. (I then got a major headache lasting days, and then focused on learning structured procrastination, and "actually resting" and "thinking without muscle tension", and now things feel pretty sustainable)

I spent the past year building out a framework, based on a vague vision. I now feel like I concretely have the necessary pieces for "recursively self improving" to be tractable, (including both explicit cognitive skills, tacit knowledge articulation/interviewing, bodily and soulful self-awareness skills, and some object level knowledge).

I'm just getting started on the Cyborg tooling part of this plan, and it's much more uncertain. But I've done some experiments with retooling my cognitive style such that I:

- think a thought
- ask "could an LLM do this?"
- if the answer is "probably?", then, have an LLM take a first stab at it, and move on to the next thought. The answer is "probably" pretty often when coding, and sometimes while brainstorming/strategizing/gathering-info.

- Eventually, (when both LLMs improve and it seems tractable for my "thoughts" to be more like "the table of contents of my thoughts."

I think using LLMs productively for alignment research seems much harder. But, it seems at least quite tractable to streamline many kinds of learning about existing topics. I have some preliminary ideas on what might be necessary to build LLM tools / finetunings that allow it to differentially help with "technical philosophy", and I'm interested in people's cruxes / guesses-of-bottlenecks that'd need to be dealt with for that to work.

(I recently asked Zac Hatfields Dodds from Anthropic: "what if you guys just actually focused on AI that was good at philosophy instead of, like, computer programming and prosaic ML?" and he said "well, there just isn't enough money in that to be a profitable self-powering engine." On one hand, that's sad. On the other hand: this suggests it might be something someone with philanthropic support could actively build towards, such that during Golden Hour, alignment researchers can be remotely competitive with capabilities researchers)

I don't know that I'll achieve the ambitious goals here but it currently feels healthy and worthwhile to try.

Some major cruxes of mine

I am quite confident I can take Level 4-6 people (i.e. me, most members of Lightcone team, Garrett Baker and George Wang) and turn them into Level 7-8 people (Oli, Wentworth)

I am not sure whether I can take Level 8-9 people and turn them into Level 10-12 people. (Eliezer, Paul Christiano).

I'm not sure whether it actually matters, to have more level 7-8 people. It currently seems to me to matter at least some.

But almost all the people I know who are smarter than me, also seem metacognitively stupid in at least some ways. Some of those ways are straightforward and low-hanging-fruit looking. Some of them seem stickier and kinda impossible to change,

but I'm currently in the mood to look at impossible things and ask "okay, but what's actually impossible about that?" and roll up my sleeves and deal with it.

I'm currently working on the lens of "can I meaningfully help Oliver, who clearly has more g than me, and is reasonably cognitively flexible, but is pretty worried about doing stuff to his cognition on-purpose without breaking something?". If the two of us can find a thing that works, I'll feel a bit of hope trying to help people like, say, Scott Garrabrant, who seems brilliant but not really able to direct his cognition on-purpose.

Current bets I'd make

(Giving a lower and upper bound for what feels plausibly right, mostly going off gut)

⚖⚖ A year from now, I'll have a metric for Lightcone velocity that the team + 2 major stakeholders agree is notably more objective than "well, we sure *feel* faster", that is streamlined enough to actually use (Raymond Arnold: 30% - 75%)

⚖⚖ Conditional on having a LC velocity metric in 6 months, in 12 months, Lightcone will be at least 1.5x what they were 6 months from now, in the 12th month (Raymond Arnold: 20% - 60%)

⚖⚖ A year from now, Lightcone + at least 3 external people will say "Lightcone sure seems substantially faster at execution, plausibly 2x faster in the past 2 months than they were a year ago" (Raymond Arnold: 20% - 60%)

⚖⚖ A year from now, Timaeus leadership will have two plans they believe in that are not Singular Learning Theory focused. (Raymond Arnold: 35% - 80%)

⚖⚖ 3 months from now, in the past week, at least 3 Lightcone team members will each have executed two Ray-descended techniques of their own initiative (which they self report as Ray being causal in) (Raymond Arnold: 20% - 65%)

The Cognitive Bootcamp Agreement

Slightly edited version of the [public agreement](#) for workshop participants.

Two major failure modes I'm worried about with rationality training:

1. The curriculum is too opinionated to fit into participant's actual life trajectories

2. The curriculum is too openended, and people don't end up learning anything specific, and I don't get data that helps me build a comprehensive theory of change.

Since the third workshop, I've settled into a format aiming to thread the needle between those. Participants are asked to read this agreement and confirm "which parts they are bought into, and which parts they are not", and then we make a call on whether the workshop is right for now.

The target audience is people who 1) have decision-making power on a large project, which tackles confusing problems; 2) are not bottlenecked on executive function; and 3) who don't have a sneaking suspicion they're at risk of burnout.

The Format

By coming to the workshop, you agree that you'll spend the whole time either:

1. aiming to work/think basically as hard as you can.
2. napping / taking a walk / etc (preferably without devices)
3. talking to me if the workshop feels off in some way.

Or, put another way, the three major TAPs of the workshop are:

1. *"This exercise feels fake"*
    a. -> come talk to me, and either we figure out a non-fake-feeling version, or something else to do we both believe in.
2. *"This exercise feels bad for me somehow"*
    a. -> also come talk to me, and either we figure out a non-fake-feeling version, or something else to do we both believe in.
3. *"This exercise feels kind of annoying and effortful"*
    a. -> *also* come talk to me, but, basically, you can take a break but when you come back you should just do the thing.

Your goal is to learn at least one new skill, starting from "it's too cumbersome to use productively", and aiming to reach "Juuuust fluent enough that you can start applying it to your day job, practicing so it becomes easy."

Beforehand, you send me your default plan for the next ~week, month or quarter – whatever the longest timescale you plan on. You'll work on improving this plan (though the main workshop goal is still "gain one new skill")

Leveling up at least one skill

I'll present ~4 skills I think are valuable for solving confusing problems, and exercises that are helpful for grinding on them.

You're welcome to pick other skills, or other exercises, that you think will help you better. But, for each session, you need to clearly explain a) why I think the original exercise was important, b) why the thing you'll do instead is better. (We'll chat until we both understand each other and feel good about it)

The Goal: Solve confusing, intractable problems

Many of the most important problems in the world are confusing, intractable, and seem impossible. They are really important to solve anyway.

But many people who try, end up in failure modes such as…

1) **Rabbit-holing,** fixating on an approach which doesn't work, or takes too long.
2) **Goodharting**, substituting an easier problem, maybe without even noticing.
3) **Despair**. It's impossible. You give up.
4) **Zombified Agency.** You mechanically execute virtuous-seeming plans, but something inside you is dead and hollow and the plan probably won't work and maybe you'll hurt yourself.

I want you to leave the workshop with at least one new skill for solving impossible problems.

The Curriculum

The default curriculum focuses on "making better plans" via…

1. **Generating multiple plans,** so you don't over-anchor on your first plan.

2. **Tracking multiple goals**, so you don't over-anchor on your first goal.
3. **Identifying your cruxes,** and confidently deciding when to pivot or persevere.
4. **Making cruxy predictions**, so over time you can credibly, calibrated trust your intuition (and, know the limits of where you cannot)
5. **Identifying how you could have thought it faster**, and generalizing takeaways from that.

An optional, higher level frame is **Fractal Strategy** – thinking of plans and goals at multiple levels, tracking how smaller plans fit into bigger plans, and when it's time to move up, down or sideways in plan/goal-space.

If you feel like you roughly have all these skills, you might still want to come to the workshop to basically "do the cognitive practice *you* believe in with good form, attending carefully to each step of the process." (I'll want to talk to you beforehand about your plan in detail)

Additional skills

The one skill I will require everyone to attempt at least twice is **generating metastrategies** (see below).

Some more specific skills that will be presented, which we can chat more about if it feels alive and useful to you:

- *Using externalized working memory,* both because it increases the complexity of problems you're capable of tackling, and because it makes it easier for me to see your thought process and offer advice.
- *Noticing metacognition*, to identify when you're in particular cognitive states so that you can learn about your mind and employ relevant skills/habits.
- *Grieving,* so you can let go of attachments that are distorting your thinking.
- *Structured procrastination and Actually Resting*, so you don't overtax your brain.
- *Deliberate OODA Looping*, so you aren't merely making plans, but building a good engine that takes in information and improves your plans.

Multiple feedback-loops

I don't have perfect feedback-loops to tell you if this workshop is working for you. So, there are four different feedback-loop types, with different tradeoffs:

1. *Predictions*
    a. Guess whether a given strategy will work, then see if you were right.
2. *Toy Exercises.*
    a. They only vaguely resemble your real problems, but you'll know for sure whether you got the right answer in two hours.
3. *Big picture planning.*
    a. You'll generate at least one new plan. You won't really *know* if it's good, but a) you'll have intuitions about it, which are at least some information. And, b) you'll make predictions about whether it'll seem worth having thought about in a year.
4. *Object-level work, in 1-hour blocks*
    a. Spend a few timeblocks doing *object level work* on your *second likeliest* plan. Each hour, you'll make conscious choices about how to spend your time and attention. And then, reflect on whether that seemed useful. (in addition to crosstraining your skills on the practical object-level, this will help make your second-likeliest plan feel more real)

The Immediate Metric: General Strategies

The fundamental skill of the workshop is "brainstorming metastrategies." That is, general strategies that work on a wide variety of problems (usually by helping you generate more "tactical" strategies appropriate to the situation)

Each session, you'll be doing a puzzle, or intellectual work. The point isn't whether you succeed or not on the puzzle. The point is whether you learn generalized strategies that either definitively helped you on the problem, or seem likely to help you in the future.

After each session, you'll:

- First, brainstorm as many general takeaways as you can.

- Then, if there were any strategies that concretely helped, you will write it on the whiteboard.
- Put a star on it if it's a strategy you've never tried before.

Each time you successfully use a strategy already listed on the whiteboard (including the first time you write it down), put your initials next to it.

The (slightly Goodharty but hopefully productive) goal of the workshop is to fill the board with strategies that get used lots of times.

There will be a prize for:

- the person who generates the most strategies that they successfully used
- the person whose strategies got the most successful uses across participants
- the person who generated the most strategies they'd never used before

The way you're supposed to tell if the workshop is helping is "are you generating strategies that help you?".

If by the end of day 1, if you haven't generated any metastrategies, we'll chat about whether we should adjust anything on day 2 for you. If by the end of day two if you haven't generated any strategies, we'll check in on whether the workshop is working for you or you are otherwise getting value out of it, and maybe call it quits or radically change what you focus on for day 3.

The Disclaimers

Prerequisites

This is not an entry level rationality workshop. It has several prerequisites:

1. **Executive function.** You can sit down and think about a confusing problem and not immediately bounce off. Your problem should be more like "when I sit down to planmake, I don't know what to do" than "I have trouble sitting down to plan," or "if I make a decision, I can't follow it through."

2. **Project Ownership.** You have control over an (at least somewhat) open-ended decisionmaking process. A primary skill at this workshop is learning when to pivot. You need to be capable of deciding when you're pivoting. When you leave the workshop, you have a project you expect to be applying the techniques to.
    a. *i.e. you get at least some leeway to set priorities at your day job, or you have a lot of slack for ambitious hobbies.*

3. **Self awareness.** You can notice when pushing yourself hard is bad for you, instead of good for you. You have at least some awareness of when things are going subtly wrong and you need to slow down.
    a. I will do my best to also help you notice this sort of thing, but there's realistically a limit to how much I can.

If you are at risk of burnout (if you've recently worked much harder than usual, or a nagging voice inside is worried about spending a weekend doing very intense thinking) I recommend you do not attend right now. You can register your interest for future workshops though.

This will be very meta

We are going to think about thinking.

We are going to apply feedback-loops to our feedback-loops.

While we're doing that, I will be thinking about thinking about you thinking about those things (You don't have to do that, just me).

Meta-level optimization is the mechanism by which, maybe, I expect people to get compounding returns on thinking, and there is a real possibility that this can be leveraged into dramatically better plans.

Well designed feedback loops are the grounding mechanism to check whether we're doing masturbatory meta, vs useful meta.

Many people find "meta" overwhelming. Many kinds of meta *are* more masturbatory than helpful. If you are feeling disoriented or annoyed or have a nagging feeling the current amount of meta won't help:

1. first, pause (if I'm the problem, say "hey Ray stop")
2. probably, go back to doing a more object level thing.
3. but, also: consider getting out a sheet of paper / google-doc and writing things down so you can actually track what's going on

Part of the point of becoming fluent in "working memory extension" is that your meta processes are much easier to understand, and you can leverage them more strategically.

(Note: I think most people should spend ~10% of their time on meta. I'm aiming to spend more like 50% because I'm particularly specializing in it)

Why should I think this works?

I am talking a pretty big game. I think it's an active ingredient for me to present content confidently, and for people to lean into a mindset where they trust it'll work, or at least is worth trying.

But an important truth is that this workshop series does not (yet) have a strong empirical track record. If the curriculum does not intuitively make sense to you, I don't think you should particularly believe in it.

I am trying to stick my neck out such that if the workshop is not working, it should be obvious (i.e. people will not be generating strategies that help).

Here are some facts about past participant experience:

- When I ran the first workshop (which I charged $200 for), I asked the 9 participants what was the most they'd have paid for the workshop. Numbers ranged from $300 to ~$2000, on average $800.
    - When I asked again six months later, John Wentworth and Ben Weinstein Raun (who both originally said they'd pay "$400") changed their number to "$2000".

        - John, because he changed his plans importantly afterwards, probably saving a month of work. They weren't sure how much credit to assign to the workshop, but it could be anywhere from $0 to $10,000, and $2000 felt about right for what they should have been willing to pay, in expectation.
        - Ben, because they gained an major insight that was still seemed important to their worldview six months later
    - Luke Stebbing (works at Elicit) originally said $1500, and upped it to $2000, because they felt they'd improved their plan significantly in a way that stuck more than they expected. (They did separately note that they hadn't successfully integrated the skills from the workshop into their day-to-day, although had some reasons to think they might make more progress on that soon).
    - Rafe originally said "$300" and changed it to "$0 to $1000 but probably like $200."
    - One $1500 person dropped to $0 because they realized they were too burned out to get value from the workshop (and maybe it hurt them)
    - One person had said $800, dropped to $700 at the six months mark.
- When I ran the second workshop, average rating was $540, or $650 if you throw out one data point from someone (who rated it $0) who came for nonstandard reasons, and I wouldn't have really expected to get much value out of it, if we'd talked a bit more.
- For the third workshop, people gave less precise answers, but then they all were happy to pay $100/month for Recursive Self Improvement club, for awhile at least.

What's up with the cost? Can I get a discount?

I charge a fair amount of money (and plan to charge more as I become more confident in the curriculum), because:

- **People thinking it's worth paying for is a crux.** If I didn't think I could ultimately generate thousands of dollars worth of value for people, I would give up on the project. Charging money makes it harder to delude myself that I'm

helping people.

- **Lightcone needs the money.** And on the margin, we prefer to get money from people we are delivering value to.

- **Filter for commitment.** I want to filter for people who actually are going to *try* to get hundreds or thousands of dollars worth of value from the workshop, and so will put more effort in.

- **Soft filter for the pre-requisites.** Because the workshop isn't a 101 workshop, I want to filter for people who have some degree of "already having your shit together", for which "can afford to consider paying serious money for a workshop" is a proxy. (I realize that will exclude some people unnecessarily. But given the other two goals, seems like the right call)

In general, if you didn't feel like you got your money's worth, I prefer to solve this by giving you some free followup coaching until you feel like you got your money's worth.

If you earnestly tried to get a thousand dollars worth of value from the workshop, including explicitly strategizing with me on how to adapt it to you, and it didn't really pay off and it seems like my general frame or skillset isn't useful to you, chat with me and I'll consider a refund.

"Didn't the last few attempts to train rationality superhumans basically fail? Whence your hope, Ray?"

I expect to at least fail in interesting new ways. The diffs I'm aware of:

Focus on selection, and training, not recruitment.

- I think CFAR veered into doing recruitment more than training. It may have made sense at the time, but nowadays it doesn't feel like a bottleneck that will distort my attentions.
- I'm explicitly targeting "people who have demonstrated at least some promise", and trying to accelerate them.
- One of my models is "people need like 15-30 skills for rationality training to hit escape velocity. We have enough people around with 10-20 skills that it's more tractable to filter for "has most pre-reqs" and get them the other 5-10

Focus on "develop better feedback-loops" as a core project goal

- CFAR didn't emphasize this much. I think Leverage did, but Leverage had some other problems.
- Generally, I aim to have a good variety. My current main feedbackloops are similar to the ones I give the workshop participants (i.e. building better toy problems, cruxy predictions), but my main immediate ones are:
    - "how much do people report the workshop being worth, 6 months later"
    - "how much are Lightcone staff opting into my coaching / facilitation"
    - "how much are workshop participants choosing to return to Recursive Self Improvement Club."

Emphasis on longterm practice over "transformational weekend."

- I think deliberate practice basically works, but does require peak cognitive hours, which are expensive. My goal here is to get people to "escape velocity" with key skills, such that they can afford to not merely "practice", but "deliberate practice" them on the job.

- I've only just started Recursive Self Improvement Club, but I plan to build that into an engine of Rationality Science. So far there are 4 people who seem pretty bought in.

We just kinda know a lot of pitfalls and are generally older and wiser.

- I think early CFAR started with a number of really competent founders. It gradually bled them away, and I don't think it got to "a basic foundational

understanding of what was necessary" until most had left. Maybe there are more pitfalls, but, I think I'm doing some standing-on-shoulders-of-predecessors.

- There's just a whole lot of knowledge in the water about cruxes, focusing, goal factoring, operationalizing predictions, etc, such that I can focus workshops on skipping ahead to new parts.

I think I've personally hit the ~20 someodd subskills necessary for Recursive Self Improvement to work, and also think I've got "tacit knowledge articulation".

And, while that doesn't mean this program could scale arbitrarily (I think it'll be bottlenecked on me for awhile), it doesn't really need to scale arbitrarily, I think if I unlocked 50-200 people's ability to tractably tackle arbitrary confusing problems with flexible metacognition, that'd be better than most other plans I can think of.

Recursive Self Improvement Club

Recursive Club is the engine that I expect to use to push the art forward. I'm not sure of the cadence but it'll probably be twice-monthly, meeting for ~5 hours.

Notable features:

- Costs at least $100, filtering for seriousness
- Starts with one exercise from an already written up set of exercises. You can add your *own* exercises to that list, but, you have to do so before the meeting. (Last time, I told people they could pick things to do based on their own needs, but then they kind of floundered around because they didn't know how to design exercises for themselves on-the-fly)
    - If you design your own, it should be formatted in a way that a person-who-is-not-you could plausibly follow the instructions.
- Then, spend an hour thinking about why you aren't omnicompetent yet and what your next bottleneck is.

- You don't have to *believe* in Rationalist Recursive Self Improvement that isn't fake, but, you gotta be okay with being vaguely socially involved with a scene about it and periodically getting social pressure about it

Longterm, some things that seem necessary:

- Members gain fluency with inventing exercises for themselves
- Members gain passable skill at tacit knowledge interviewing
- Members are tracking their bottlenecks and skill trees they care about
- Members have outside friends who they kinda check in with to make sure they / we / I-in-particular aren't going crazy

What would I do if not this?

1. Cause the AI-risk community to make a serious chunk-of-group-epistemic-progress on the "Is Alignment Very Hard?" debate.
    a. I've chatting with Zac Hattsfield Dodd from Anthropic about this and he is pretty up for contributing seed comments in a post asking people to identify what would change their own mind on "whether AI requires 10+ years of serial research", which he agreed was a crux for whether Anthropic should be doing something different.
    b. I'd hope to get starter comments from people from a variety of perspectives lined up before publishing the post.
    c. (This seems important both for object level value a la "lots of people are either pursuing bad plans at best, or working at cross-purposes at worst", and also because the fact that we haven't resolved this, even among clearly well-meaning and smart people, feels like a bottleneck for the rest of the world actually trusting the x-risk community, which feels like a bottleneck for civilization being generally sane/competent in some ways)

2. Build some exobrain tools focused on getting me information from around-the-world, which alert me to news, existing research, tools or people that I don't know about, which might change my plans.
    a. (I feel like the "observe" part of my OODA loop is my weakest link)
    b. (if this is worth doing, it's probably worth doing soon, so it can start affecting my plans sooner)

3. Focus on building LLM-exobrain tools for John Wentworth and Chase Deneke in particular, since they're just here at Lighthaven all the time and both seem to be working on plausibly quite important stuff. (Or, do more of a search for people who make sense to try this for)

4. Stop trying to uplift anyone else along with me and just focusing on accelerating my own competence, and tackle some difficult projects that'd previously have been beyond me, and see how that goes.

5. Figure out if I should get over myself and move to DC and build some kind of sanity-inducing community over there.

Questions for Nate

(probably easiest to chat about these in-person. Probably best to do after you've at least skimmed the rest of the doc, but)

What, if any, internal rationality training did MIRI do, that isn't part of "the public CFAR canon?"

What seem like the essential skills that an alignment researcher needs to tackle "the hard parts" of the alignment problem?

You and Eliezer have tried mentoring various people, and (I think?) it basically hasn't worked according-to-you. What actually happened, and why didn't it work?

- Tried Vision Transfer
    - outsourcing meaningful chunks of cognitive labor of the hard parts to another person
- Tried
    - tried on big abstract alignment problems
    - …small concrete math domains, lots more literature and visible feedback
    - a lot of regular interaction and aesthetic discussions
    - people have their own sense of aesthetics, chat and sync sometimes and let the develop their own variant
- None of that translated into
- Some people
    - Rafe
        - more contact and attempt to do aesthetic transfer
    - Jesse Liptrapp
        - concrete math stuff
            - jesse having own directions
        - jesse came closest to having insights where I'm like "that's a small fraction of labor I would have wanted to do."
    - Jason Gross
        - concrete math stuff
            - more talking details
- People who maybe could 100x and then we'd be fine:
    - Nate
    - Benya
    - Eliezer
- "Maybes"
    - Mark Xu
    - JohnW

How much has Nate X'd himself

- About a year of gaining
    - working with other people who working on the problems

What would be more hopeful for Ray or someone

- People who could be visionary leaders of brain uploading or cognitive enhancement

What's the broken part?

- research vision
    -

Alice

- Trying to transfer whole alignment intuitions

Bob

- More object level, do aesthetically large projects

Rafe

- Have better mathematical intuitions.
- sometimes people have intuition, some aesthetic hard-to-verbalize sense that there's something important in some direction.
    - when they have it, it pulls them in a direction that is accurate but personal
    - if people are lucky enough to have some of these, they can choose whether to p
    - mathematical acumen

        -
    - raw intelligence
-

Jesse

-

Jason

-

Vivek

- me trying to get them to have ideas of their own
    - in large part because they weren't able to have good ideas of their own

A place people fall down is:

- they actually don't think for themselves
- Vivek was like "X can't possibly be so, because it was people would have noticed."

Jeremy Gillan (on Vivek team) was best at not

John W

- very sloppy
- very prone to thinking he had a result when he didn't

Nate:

- currently writing a book with Eliezer
- Normal arguments, for broad

Nate

- for each clone:
    - lobbying politicians
    - go help Benja with Benja research
        -
    - agent foundations
        -
    - what are the possible last-ditch efforts where these things are somehow friendly somehow
        - constructing a giant dataset that has filtered
        - could you get an AI that thinks it's friendly, long enough to help

Understanding intelligence

- agent foundations
- studying GPTs to the point where we understand what
-

If you totally believe

For AF to work in real life, a lot of the goal is, a lot of the goal in 2014 was "cause there to be a healthy field here." Nate and Benja and Jessica are gonna solve this all themselves.

First step is for Rafe

What currently seem like the bottlenecks for extracting useful cognitive labor out of LLMs, to assist a researcher targeting the hard part(s) of the alignment problem?

What do you think an older, wiser and/or more cynical version of me would know-in-my-heart, that I'd wish I'd internalized now?

Who are the people you think are the best at "Applied Metacognition?" and why?

Alternately, who are the people you know of who seem to be capable of end-to-end Human Recursive Self Improvement?

What are two things I should strongly consider doing, instead of this agenda?

It'd be pretty embarrassing if I were the "have 3+ plans" guy without having 3+ plans I believed in. Some alternative goals I might focus on at the 6-24 month level. These are noticeably more vague, and a TODO is to flesh them out more such that I have real alternatives.