Best of LessWrong: December 2014

Best of LessWrong: December 2014

We Haven't Uploaded Worms

> In theory you can upload someone's mind onto a computer, allowing them to live forever as a digital form of consciousness, just like in the Johnny Depp film Transcendence.
>
> But it's not just science fiction. Sure, scientists aren't anywhere near close to achieving such feat with humans (and even if they could, the ethics would be pretty fraught), but now an international team of researchers have managed to do just that with the roundworm Caenorhabditis elegans.
> —Science Alert

Uploading an animal, even one as simple as *c. elegans* would be very impressive. Unfortunately, we're not there yet. What the people working on Open Worm have done instead is to build a working robot based on the *c. elegans* and show that it can do some things that the worm can do.

The *c. elegans* nematode has only 302 neurons, and each nematode has the same fixed pattern. We've known this pattern, or connectome, since 1986. [1] In a simple model, each neuron has a threshold and will fire if the weighted sum of its inputs is greater than that threshold. Which means knowing the connections isn't enough: we also need to know the weights and thresholds. Unfortunately, we haven't figured out a way to read these values off of real worms. Suzuki et. al. (2005) [2] ran a genetic algorithm to learn values for these parameters that would give a somewhat realistic worm and showed various wormlike behaviors in software. The recent stories about the Open Worm project have been for them doing something similar in hardware. [3]

To see why this isn't enough, consider that nematodes are capable of learning. Sasakura and Mori (2013) [5] provide a reasonable overview. For example, nematodes can learn that a certain temperature indicates food, and then seek out that temperature. They don't do this by growing new neurons or connections, they have to be updating their connection weights. All the existing worm simulations treat weights as fixed, which means they can't learn. They also don't read weights off of any individual worm, which means we can't talk about any specific worm as being uploaded.

If this doesn't count as uploading a worm, however, what would? Consider an experiment where someone trains one group of worms to respond to stimulus one way and another group to respond the other way. Both groups are then scanned and simulated on the computer. If the simulated worms responded to simulated stimulus the same way their physical versions had, that would be good progress. Additionally you would want to demonstrate that similar learning was possible in the simulated environment.

(In a 2011 post on what progress with nematodes might tell us about uploading humans I looked at some of this research before. Since then not much has changed with nematode simulation. Moore's law looks to be doing much worse in 2014 than it did in 2011, however, which makes the prospects for whole brain emulation substantially worse.)

I also posted this on my blog.

[1] The Structure of the Nervous System of the Nematode Caenorhabditis elegans, White et. al. (1986).

[2] A Model of Motor Control of the Nematode C. Elegans With Neuronal Circuits, Suzuki et. al. (2005).

[3] It looks like instead of learning weights Busbice just set them all to +1 (excitatory) and -1 (inhibitory). It's not clear to me how they knew which connections were which; my best guess is that they're using the "what happens to work" details from [2]. Their full writeup is [4].

[4] The Robotic Worm, Busbice (2014).

[5] Behavioral Plasticity, Learning, and Memory in C. Elegans, Sasakura and Mori (2013).

CFAR in 2014: Continuing to climb out of the startup pit, heading toward a full prototype

Summary: We outline CFAR’s purpose, our history in 2014, and our plans heading into 2015.

One of the reasons we’re publishing this review now is that we’ve just launched our annual matching fundraiser, and want to provide the information our prospective donors need for deciding. **This is the best time of year to decide to donate to CFAR. Donations up to $120k will be matched until January 31.**[1]

To briefly preview: For the first three years of our existence, CFAR mostly focused on getting going. We followed the standard recommendation to build a ‘minimum viable product’, the CFAR workshops, that could test our ideas and generate some revenue. Coming into 2013, we had a workshop that people liked (9.3 average rating on “Are you glad you came?”; a more recent random survey showed 9.6 average rating on the same question 6-24 months later), which helped keep the lights on and gave us articulate, skeptical, serious learners to iterate on. At the same time, the workshops are not everything we would want in a CFAR prototype; it feels like the current core workshop does not stress-test most of our hopes for what CFAR can eventually do. The premise of CFAR is that we should be able to apply the modern understanding of cognition to improve people’s ability to (1) figure out the truth (2) be strategically effective (3) do good in the world. We have dreams of scaling up some particular kinds of sanity. Our next goal is to build the minimum strategic product that more directly justifies CFAR’s claim to be an effective altruist project.[2]

Highlights from 2014

Our brand perception improved significantly in 2014, which matters because it leads to companies being willing to pay for workshop attendance. We were covered in Fast Company -- twice -- the Wall Street Journal, and The Reasoner. Other mentions include Forbes, Big Think, Boing Boing, and Lifehacker. We’ve also had some interest in potential training for tech companies.

Our curriculum is gaining a second tier in the form of alumni workshops. We tried 4 experimental alumni workshops, 3 of which went well enough to be worth iterating:

- The Hamming Question: “What are the most important problems in your life, and why aren’t you working on them?” This 2.5-day workshop was extremely well received, and gave rise to a new unit for our introductory workshop.

- Assisting Others[3]: A two-weekend (training, then practicum) workshop investigating the close link between helping others debug their problems, and better debugging your own problems. We ran a version of this in the Bay Area that worked, and an abridged version in the UK that didn't. (This was our fault. We're sorry.)
- Attention Workshop: A 2.5-day workshop on clearing mental space. This failed and taught us some important points about what doesn't work.
- Epistemic Rationality for Effective Altruists: A standalone 2.5-day workshop on applying techniques from the introductory workshop to factual questions, especially those related to effective altruism. (More on this below.) The attendees from this and the Hamming workshop spontaneously organized recurring meetups for themselves.

Our alumni community continues to grow. There are now 550 CFAR alumni, counting 90 from SPARC. It's a high-initiative group. Startups by CFAR alumni include: Apptimize; Bellroy; Beeminder; Complice; Code Combat; Draftable; MealSquares; OhmData; Praxamed; Vesparum; Teleport; Watu; Wave; ZeroCater.[4] There is a highly active mailing list with over 400 members, and over 600 conversation threads, over 30 of which were active in the last month. We also ran our first-ever alumni reunion, and started a weekly alumni dojo. This enabled further curricular experimentation, and allowed alumni ideas and experiences to feed into curricular design.

SPARC happened again, with more-honed curriculum and nearly twice as many students.

Basic operations improved substantially. We'll say more on this in section 2.

Iteration on the flagship workshop continues. We'll say more on this (including details of what we learned, and what remains puzzling) in section 3.

Improving operations

The two driving themes of CFAR during 2014 were making our operations more stable and sustainable, and our successful struggle to pull our introductory workshop out of a local optimum and get back on track toward something that is more like a 'full prototype' of the CFAR concept.

At the end of 2013, we had negative $30,000 and had borrowed money to make payroll, placing us in the 'very early stage, struggling startup' phase. Almost all of our regular operations, such as scheduling interviews for workshop admissions, were being done by hand. Much of our real progress in 2014 consisted of making things run smoothly and getting past the phase where treading water requires so many weekly hours that nobody has time for anything else. Organizational capital is real, and we had to learn the habit of setting aside time and effort for accumulating it. (In retrospect, we were around a year too slow to enter this phase, although in the very early days it was probably correct to be building everything to throw away.)

A few of the less completely standard lessons we think we learned are as follows:

- Rank-order busyness, especially if you're passing up organizational-capital improvement tasks. Think "This is one of the 3 busiest weekends of the year" and not "I'm too busy to do it right now." This says how large a hit you get from allowing “important but not urgent” to be postponed during times which are at least that busy, and it forces calibration.

- Even in crunch times, take moments to update. (E.g., do one-sentence journal entries about what just happened / ideas for improvement after each Skype call.) The crunchiest moments are often also the most important to optimize, and even a single sentence of thought can give you a lot of the value from continuing to optimize.
- Use arithmetic to estimate the time/money/staff cost of continuing to do Y the usual way, versus optimizing it. If the arithmetic indicates 10X or more savings, do it even if it requires some up-front cost. (No really, actually do the arithmetic.)

We also learned a large number of other standard lessons. As of the end of 2014, we think that basic processes at CFAR have improved substantially. We have several months of runway in the bank account - our finances are still precarious, but at least not negative, and we think they’re on an improving path. Our workshop interviews and follow-up sessions have an online interface for scheduling instead of being done by hand (which frees a rather surprising amount of energy). The workshop instructors are almost entirely not doing workshop ops. Accounting has been streamlined. The office has nutritious food easily available, without the need to quit working when one gets hungry.

CFAR feels like it is out of the very-early-startup stage, and able to start focusing on things other than just staying afloat. We feel sufficiently non-overwhelmed that we can take the highest-value opportunities we run into, rather than having all staff members overcommitted at all times. We have a clearer sense of what CFAR is trying to do; of what our internal decision-making structure is; of what each of our roles is; of the value of building good institutions for recording our heuristic updates; etc. And we have will, momentum, and knowledge with which to continue improving our organizational capital over 2015.

Attempts to go beyond the current workshop and toward the ‘full prototype’ of CFAR: our experience in 2014 and plans for 2015

Where are we spending the dividends from that organizational capital? More ambitious curriculum. Specifically, a "full prototype" of the CFAR aim.

Recall that the premise of CFAR is that we should be able to apply the modern understanding of cognition to improve people’s ability to (1) figure out the truth; (2) be strategically effective; and (3) do good in the world. By a “prototype”, or “minimum strategic product”, we mean a product that actually demonstrates that the above goal is viable (and, thus, that more directly justifies CFAR's claim to be an effective altruist project). For CFAR, this will probably require meaningfully boosting some fraction of participants along all three axes (epistemic rationality; real-world competence; and tendency to do good in the world). [5]

So that's our target for 2015. In the rest of this section, we'll talk about what CFAR did during 2014, go into greater detail on our attempt to build a curriculum for epistemic rationality, and describe our 2015 goals in more detail.

One of the future premises of CFAR is that we can eventually apply the full scientific method to the problem of constructing a rationality curriculum (by measuring variations, counting things, re-testing, etc.) -- we aim to eventually be an evidence-based organization. In our present state this continues to be a lot harder than we would like; and our 2014 workshop, for example, was done via crude "what do you feel you learnt?" surveys and our own gut impressions. The sort of randomized trial we ran in 2012 is extremely expensive for us because it requires randomly not admitting workshop attendees, and we don't presently have good-enough outcome metrics to justify that expense. Life outcomes, which we see as a gold standard, are big noisy variables with many contributing factors - there's a lot that adds to or subtracts from your salary besides having attended a CFAR workshop, which means that the randomized tests we can afford to run on life outcomes are underpowered. Testing later ability to perform specific skills doesn't seem to stress-test the core premise in the same way. In 2014 we continued to track correlational data and did more detailed random followup surveys, but this is just enough to keep such analyses in the set of things we regularly do, and remind ourselves that we are supposed to be doing better science later.

At the start of 2014, we thought our workshops had reached a point of decent order, and we were continuing to tweak them. Partway through 2014 we realized we had reached a local optimum and become stuck (well short of a full prototype / minimum strategic product). So then we smashed everything with a hammer and tried:

- 4 different advanced workshops for alumni:
  - An epistemic rationality workshop for effective altruist alumni;
  - An alumnus workshop on focusing attention (failed);
  - An alumnus workshop on the Hamming Question, "What are your most important life problems? Why aren't you solving them?"
  - 2 attempts at an alumnus workshop on how to do 1-on-1 teaching / assistance of cognitive skills (first succeeded, second failed; our fault).
- A 1.5-day version of the introductory workshop;
- A workshop with only 10 participants with the entire class taught in a single room (extremely popular, but not yet scalable);
- Shorter modules breaking up the 60-minute-unit default;
- An unconference-style format for the 2014 alumni reunion.

These experiments ended up feeding back into the flagship workshop, and we think we're now out of the local optimum and making progress again.

Epistemic rationality curriculum

In CFAR's earliest days, we thought epistemic rationality (figuring out the answers to factual questions) was the main thing we were supposed to teach, and we took some long-suffering volunteers and started testing units on them. Then it turned out that while all of our material was pretty terrible, the epistemic rationality parts were even more terrible compared to the rest of it.

At first our model was that epistemic rationality was hard and we needed to be better teachers, so we set out to learn general teaching skills. People began to visibly enjoy many of our units. But not the units we thought of as "epistemic rationality". They still visibly suffered through those.

We started to talk about "the curse of epistemic rationality", and it made us worry about whether it would be worth having a CFAR if we couldn't resolve it somehow. Figuring out the answers to factual questions, the sort of subject matter that appears in the Sequences, the kind of work that we think of scientists as carrying out, felt to us like it was central to the spirit of rationality. We had a sense (and still do) that if all we could do was teach people how to set up trigger-action systems for remembering to lock their house doors, or even turn an ugh-y feeling of needing to do a job search into a series of concrete actions, this still wouldn't be making much progress on sanity-requiring challenges over the next decades. We were worried it wouldn't contribute strategic potential to effective altruism.

So we kept the most essential-feeling epistemic rationality units in the workshop even despite participants' lowish unit-ratings, and despite our own feeling that those units weren't "clicking', and we thought: “Maybe, if we have workshops full of units that people like, we can just make them sit through some units that they don’t like as much, and get people to learn epistemic rationality that way”. The “didn’t like” part was painful no matter what story we stuck on it. We rewrote the Bayes unit from scratch more or less every workshop. All of our “epistemic rationality” units changed radically every month.

One ray of light appeared in mid-2013 with the Inner Simulator unit, which included techniques about imagining future situations to see how surprised you felt by them, and using this to determine whether your Inner Simulator really strongly expected a new hire to work out or whether you are in fact certain that your project will be done by Thursday. This was something we considered to be an "epistemic rationality" unit at the time, and it worked, in the sense that it (a) set up concepts that fed into our other units, (b) seemed to actually convey some useful skills that people noticed they were learning, and (c) people didn't hate it.

(And it didn't feel like we were just trying to smuggle it in from ulterior motives about skills we thought effective altruists ought to have, but that we were actually patching concrete problems.)

A miracle had appeared! We ignored it and kept rewriting all the other "epistemic rationality" units every month.

But a lesson that we only understood later started to seep in. We started thinking of some of our other units as having epistemic rationality components in them -- and this in turn changed the way we practiced, and taught, the other techniques.

The sea change that occurred in our thinking might be summarized as the shift from, "Epistemic rationality is about whole units that are about answering factual questions" to there being a truth element that appears in many skills, a point where you would like your System 1 or System 2 to see some particular fact as true, or figure out what is true, or resolve an argument about what will happen next.

- We used to think of Comfort Zone Expansion[6] as being about desensitization. We would today think of it as being about, for example, correcting your System 1's anticipation of what happens when you talk to strangers.

- We used to think of Urge Propagation[6] as being about applying behaviorist conditioning techniques to yourself. Today we teach a very different technique under the same name; a technique that is about dialoging with your affective brain until system 1 and system 2 acquire a common causal model of whether task X *will in fact* help with the things you most care about.
- We thought of Turbocharging[6] as being about instrumental techniques for acquiring skills quickly through practice. Today we would also frame it as, "Suppose you didn't know you were supposed to be 'Learning Spanish'. What would an outside-ish view say about what skill you might be practicing? Is it filling in blank lines in workbooks?"
- We were quite cheered when we tried entirely eliminating the Bayes unit and found that we could identify a dependency in other, clearly practical, units that wanted to call on the ability to look for evidence or identify evidence.
- Our Focused Grit and Hard Decisions units are entirely "epistemic" -- they are straight out just about acquiring more accurate models of the world. But they don't feel like the old "curse of epistemic rationality" units, because they begin with an actual felt System 1 need ("what shall I do when I graduate?" or similar), and they stay in contact with System 1's reasoning process all the way through.

When we were organizing the UK workshop at the end of 2014, there was a moment where we had the sudden realization, "Hey, maybe almost all of our curriculum is secretly epistemic rationality and we can organize it into 'Epistemic Rationality for the Planning Brain' on day 1 and 'Epistemic Rationality for the Affective Brain' on day 2, and this makes our curriculum so much denser that we'll have room for the Hamming Question on day 3." This didn't work as well in practice as it did in our heads (though it still went over okay) but we think this just means that the process of our digesting this insight is ongoing.

We have hopes of making a lot of progress here in 2015. It feels like we're back on track to teaching epistemic rationality - in ways where it's forced by need to usefully tackle life problems, not because we tacked it on. And this in turn feels like we're back on track toward teaching that important thing we wanted to teach, the one with strategic implications containing most of CFAR's expected future value.

(And the units we think of as "epistemic" no longer get rated lower than all our other units; and our alumni workshop on Epistemic Rationality for Effective Altruists went over very well and does seem to have helped validate the propositions that "People who care strongly about EA's factual questions are good audiences for what we think of as relevant epistemic skills" and "Having learned CFAR basics actually does help for learning more abstract epistemic rationality later".)

Goals for 2015

In 2015, we intend to keep building organizational capital, and use those dividends to keep pushing on the epistemic rationality curriculum, and pushing toward the minimum strategic project that stress-tests CFAR's core value propositions. We've also set the following concrete goals[7]:

- Find some way to track a metric for 'How likely we think this person is to end up being strategically useful to the world', even if it's extremely crude.[8]
- Actually start tracking it, even if internally, subjectively, and terribly.
- Try to boost alumni scores on the three components of "Figure out true things", "Be effective" and "Do-gooding" (from our extremely crude measure).

- Cause 30 new people to become engaged in high-impact do-gooding in some interesting way, including 10+ with outside high status and no previous involvement with EA.
- Cause 10 high-impact do-gooder alumni to say that, because of interacting with CFAR, they became much more skilled/effective/well-targeted on strategically important things. Have this also be plausible to their coworkers.

Nuts, Bolts, and Financial Details

Total expenditures
Our total expenditures in 2014 came up about $840k. This number includes about $330k of non-staff direct workshop costs (housing, food, etc.), which is offset for the associated workshop revenue; if one excludes this number, our total expenditures in 2014 came to about $510k.

Basic operating expenses
Our basic operating expenses from 2014 were fairly similar to 2013: a total of about $42k/month, outside-view:

- $5.3k/month for office rent;
- $30k/month for salaries (includes tax, health insurance, and contractors; our full-time people are still paid $3.5k/month);
- $7k/month for total other non-workshop costs (flights and fees to attend others' trainings; office groceries; storage unit, software subscriptions; ...)

Flagship Workshops
We ran 9 workshops in 2014, which generated about $435k in revenue, but also $210k in non-staff costs (mostly food and housing for workshop participants), for a total net of about $230k in additional money (or $25k/workshop in additional money), ignoring staff cost.

Per workshop staff time-cost is significantly lower than it was (counting sales, pre-working prep, instruction, and follow-ups) -- perhaps 100 person-days per workshop going forward, compared against perhaps 180 person-days per workshop in 2013. (We aim to decrease this further in 2014 while maintaining or increasing quality.)

Per workshop net revenue is on the other hand roughly similar to 2013; this was based on an intentional effort to move staff time away from short-term sales toward investment in longer-term press funnel, curriculum development (e.g., the alumni events), and other shifts to our *longer-term* significance.

Alumni reunion, alumni workshops, alumni dojo...
We ran an alumni reunion, 4 alumni workshops, and a continuing alumni dojo. We intentionally kept the cost of these low to participants, and sliding-scale, so as to help build the community that can take the art forward.
Detail:

- Alumni reunion: $34k income; $38k non-staff costs (for ~100 participants)
- Hamming: $3.6k revenue; $3k non-staff costs
- Assisting thinking: $2.1k revenue; $3.2k non-staff costs
- Attention: $3.3k revenue; $2.7k non-staff costs
- Epistemic Rationality for Effective Altruists: $5k revenue; $3k costs

- Dojo: free.

We also ran a 1.5-day beta workshop for beginners:

- “A taste of rationality”: $5k revenue; $2.6k non-staff costs.

SPARC
SPARC 2014’s non-staff costs came to $62k, and were covered by Dropbox, Quixey, and MIRI (although, as with our other programs, considerable CFAR staff time also went into SPARC).

Balance sheet
CFAR has about $130k, going into 2015. (The $30k short-term loan we took last year was repaid as scheduled, following last year's fundraising drive.)

Summary
CFAR is more financially stable than it was a year ago but remains dependent on donation to make ends meet, and still more dependent on donation if it is to e.g. outsource the accounting, to further streamline the per-workshop staff time-costs, and to put actual quality focus into developing the epistemic rationality and do-gooding impacts.

The big picture and how you can help

CFAR seems to many of us to be among the efforts most worth investing in. This isn’t because our present workshops are all that great. Rather, it is because, in terms of “saving throws” one can buy for a humanity that may be navigating tricky situations in an unknown future, improvements to thinking skill seem to be one of the strongest and most robust. And we suspect that CFAR is a promising kernel from which to help with that effort.

As noted, we aim in 2015 to get all the way to a “full prototype” -- a point from which we are actually visibly helping in the aimed-for way. This will be a tricky spot to get to. Our experience slowly coming to grips with epistemic rationality is probably more rule than the exception, and I suspect we’ll run into a number of curve balls on path to the prototype.

But with your help -- donations are at this stage critical to being able to put serious focused effort into building the prototype, instead of being terribly distracted staying alive -- I suspect that we can put in the requisite focus, and can have the prototype in hand by the end of 2015.

...

Besides donations, we are actually in a good position now use your advice, your experience, and your thoughts on how to navigate CFAR's remaining gaps; we have enough space to take a breath and think strategically.

We're hoping 2015 will also be a year when CFAR alumni and supporters scale up their connections and their ambitions, launching more startups and other projects. Please keep in touch if you do this; we’d like our curriculum-generation process to continue to connect to live problems.

A very strong way to help, also, is to come to a workshop, and to send your friends there. It keeps CFAR going, we always want there to be more CFAR alumni, and it might even help with that quest. (The data strongly indicates that your friends will thank you for getting them to come… and will do so even more 6 months later!)

And do please donate to the Winter 2014 fundraising drive!

---

[1] That is: by giving up a dollar, you can, given some simplifications, cause CFAR to gain two dollars. Much thanks to Peter McCluskey, Jesse Liptrap, Nick Tarleton, Stephanie Zolayvar, Arram Sabeti, Liron Shapira, Ben Hoskin, Eric Rogstad, Matt Graves, Alyssa Vance, Topher Hallquist, and John Clasby for together putting up $120k in matching funds.

[2] This post is a collaborative effort by many at CFAR.

[3] The title we ran it under was "TA training", but the name desperately needs revision.

[4] This is missing several I can almost-recall and probably several others I can’t; please PM me if you remember one I missed. Many of the startups on this list have multiple founders who are CFAR alum. Omitted from this list are startups that were completed before the alumni met us, e.g. Skype; we included however startups that were founded before folks met us and carried on after they became alumni (even when we had no causal impact on the startups). Also of note is that many CFAR alumni are in founding or executive positions at EA-associated non-profits, including CEA, CSER, FLI, Leverage, and MIRI. One reason we're happy about this is that it means that the curriculum we're developing is being developed in concert with people who are trying to really actually accomplish hard goals, and who are therefore wanting more from techniques than just "does this sound cool".

[5] Ideally, such a prototype might accomplish increases in (1), (2), and (3) in a manner that felt like facets of a single art, or that all drew upon a common base of simpler cognitive skills (such as subskills for getting accurate beliefs into system 1, for navigating internal disagreement, or for overcoming learned helplessness). A “prototype” would thus also be a product that, when we apply local optimization on it, takes us to curricula that are strategically important to the world -- rather than, say, taking us to well-honed “feel inspired about your life” workshops, or something).

Relative to this ideal, the current curriculum seems to in fact accomplish some of (2), for all that we don't have RCTs yet; but it is less successful at (1) and (3). (We'd like, eventually, to scale up (2) as well.) However, we suspect the curriculum contains seeds toward an art that can succeed at (1) and (3); and we aim to demonstrate this in 2015.

[6] Apologies for the jargon. It is probably about time we wrote up a glossary; but we don't have one yet. If you care, you can pick up some of the vocabulary from our sample workshop schedule.

[7] This isn’t the detailed tactical plan; we’ll need one of those separately, and we have a partial version that this margin was too small to contain; it’s meant to be a listing of how you and we can tell whether we won, at the end of 2015.

[8] The Apgar score for assessing newborn health is inspiring, here; if you've not seen it before, and you're wondering how one could possibly come up with a metric, you might glance at its wikipedia page. Basically, instead of coming up with a single 0 to 10 newborn health scale, Dr. Apgar chose 5 simpler components (newborn color; newborn heart rate; etc.), came up with *very* simple "0 to 2" measures for these, and then added.

State of the Solstice 2014

This'll be the first of a collection of posts about the growing Secular Solstice. This post gives an overview of what happened this year. Future posts will explore what types of Solstice content resonates with which people, what I've learned about how Less Wrong culture intersects with other cultures, and updates I've made about ritual as it relates to individuals as well as movement building.

For the past three years, I've been spending the last several months of each year frantically writing songs, figuring out logistics, and promoting the New York Winter Solstice celebration for the Rationality and Secular communities in NYC.

This year... well, I did that too. But I also finally got to go a Solstice that I *wasn't* responsible for. I went to the Bay Area on December 13th, traveled straight from the airport to the dress rehearsal...

...and I found a community coming together to create something meaningful. I walked into the hall and found some 30 or so people, with some stringing together lights, some people tying decorations around candles, a choir singing together... it felt very much like a genuine holiday coming together in an organic fashion.

(There was some squabbling about how to best perform particular songs... but it felt *very* much to me like *real holiday squabbling,* whenever a family of creative people with strong opinions on things get together, and I found it surprisingly heartwarming)

This year there were four large Solstices in the US and one small but intense 3-day event in Leipzig. Each of them had a somewhat different audience and a different focus. Each was put on by local communities who I encouraged to put their own spin on it, bringing a mix of new and "traditional" songs. I've gotten some interesting feedback on which parts of Less Wrong culture resonate with which people.

Solstices that I actively collaborated with and/or consulted on include:

Bay Area - The most explicitly Less Wrong-y and transhumanist Solstice this year. Approximately 130 attendees. I suspect had the largest choir leading songs. About half of the audience seemed to be from what I'd consider the collective Rationality community (i.e. EA/Less Wrong/CFAR) and half were "friends of friends" who were less Less Wrong-y.

> (If you went to the Bay Solstice, you can fill out the anonymous feedback form here. If you don't have much time, at least answering the first couple questions would be helpful)

Seattle - Run by a mix of Less Wrong and EA types. It was a shorter ceremony but included additional activities like improv games, Non-Violent Communication workshops and other ways to break the ice and help people bond. There between 50-60 attendees.

> (Feedback form for Seattle)

San Diego - This was put on by Sunday Assembly and the local Coalition of Reason (a collection of secular/humanist/atheist groups), with no Less Wrong connection at all. They also had fewer songs, more stories and other group activities. This one had about 100 adult attendees and about 20 children. (Probably the most family friendly of the bunch).

Leipzig - This was a three day workshop, with around 20 people who worked collaboratively to design and run a ritual together, along with some highlight songs from the "traditional" Brighter Than Today program.

New York - I ran this personally, co-sponsored by Sunday Assembly and Ethical Culture. My goal with the event was to highlight important ideas I've learned from Less Wrong and Effective Altruism, framed in the language of the general secular movement. I also wanted to address particular issues affecting the New York Rationality community. We had 180 attendees, about half from the Rationality/EA communities and half from various secular communities.

> (If you went to the NYC Solstice, you can fill out the anonymous feedback form here. Again, if you don't have much time, it's still helpful if you answer the first couple questions)

Washington Post Article

I was put in touch with a reporter who was excited to cover the Solstice's growth as secular holiday. She ended up attending the one in San Francisco. (We talked beforehand about how the Bay Area Solstice would be put on by a local community with a lot of Silicon Valley tech-entrepreneurship-types, how it'd likely have more of a focus on technology, and that they'd be trying out some sillier songs that I wasn't sure would work at a more mainstream event. She noted that she would focus on covering the growth of the holiday as a whole rather than focusing on the particular execution at the Bay Area event).

She wrote an article for the Religious News Wire, which was picked up by the Washington Post website among other places. The article was extremely positive (although it mistakenly attributes some speeches to me which were actually given by locals). I mentioned that to her and she edited the original, but the Washington Post had already picked up the earlier version.

A brief snippet:

> “We live in a world beyond the reach of God,” one of the service’s many readers said as 130 or so people gathered huddled over white candles in glass votives at Humanist Hall — a purple-painted house near downtown Oakland. “It is a hard universe. If we want to build a softer universe we will have to do it ourselves.” As a choir broke into “Here Comes the Sun,” an inscription painted on the wall beamed down upon the gathered, “The world is my country, to do good is my religion.”

Building a Better Solstice

I deliberately encourage experimentation with the format - real ritual evolves, and if we want *great* ritual (in particular, great ritual-for-your-particular-local-community, as opposed to great ritual-for-the-NYC-crowd that you're trying to emulate), we want to iterate faster, see what resonates with people, and let the less interesting variations die out.

This year I'm working to ensure each local solstice does a feedback form and sends it out as soon as possible. I want to get a sense of what things resonate (or don't) fairly reliably across the world and what things happened to work for particular groups or small sample sizes.

There's an obvious problem that a) the most likely people to fill out the survey are people who like the event, b) the second most likely people to fill out the survey are people who hate the event. If you felt the event was "meh" (and perhaps don't want to fill out an entire survey because you don't care that much), it'd still be helpful if you at least briefly answered the first couple questions, so we have a broader sense of what works.

(And again, we definitely want negative feedback, in particular from people who want *something* similar to the Solstice but didn't like the execution of it)

One thing that's clear is that transhumanist and LW content is polarizing, even within the Less Wrong crowd - some people really love it and it makes them feel connected and inspired. Others find it very offputting. It's possible that the best solution is to have multiple events that cater to different people. But I've also found it's very possible (albeit harder - it took me a couple years of practice) to make an event that works on multiple levels, resonating with the mainstream secular community with Easter eggs that are funny/sad/inspiring to people in the LW or Transhuman spheres.

Another thing that's perhaps more surprising: people who don't know what Less Wrong or Transhumanism are *at all* tend to be perfectly fine with Less Wrong and a lot of

Transhumanist content and can be pretty oblivious to even fairly direct anti-death messages. It's the people who *know* about the memesphere and either don't like it or are worried about their friends not liking it that are most uncomfortable.

What makes a *Secular* Solstice™?

So, if experimentation is a core element of the whole endeavor, what makes something a Secular Solstice as opposed to a solstice-that-is-secular? Or any other non-theistic holiday? The previous attempt at a humanist holiday, *HumanLight*, hasn't really caught on, and I think that's largely because it's deliberately flexible - you can celebrate it on whatever day you want, with whatever traditions you want, so long as it ties in with humanity-as-a-force-for-good.

Being *too* flexible or generic makes for a watered-down experience that nobody especially likes, or a collection of random experiences that don't have much in common with each other.

So my answer is this:

The core element of the Brighter Than Today Solstice is the emotional arc. It begins fun and upbeat. It turns somber, then sad, before turning uplifting and inspirational. It should lead people from light to darkness to light again, both literally and metaphorically.

The single most important image people should imagine, when they are visualizing the solstice, is the Candlelit Story and the Moment of Darkness - a moment when all but one candle has been extinguished, and a story is told about the hardships we've faced and the hardships yet to come. The story should end giving people reason to hope, without resorting to comforting falsehoods.

Then that candle is extinguished, and people sit in the darkness for a minute before the lights are rekindled and hope returns in earnest.

The particular songs I've written and found are not essential, nor is even the idea of *music* itself necessarily. (Although I highly recommend it, and I highly recommend finding songs that fill similar purposes. You might not be a fan of the actual song Brighter Than Today, but it's very helpful to have some kind of emotionally uplifting piece that's rooted in an evidence-based worldview that lifts you out of the darkness).

Some people actively dislike music or group singing, but still like the stories that take them through that arc. Some people like "preachy" content that gives people a call to action, and some people hate it and prefer more casual personal stories.

I'm not sure how this holiday will evolve to best meet the needs of the rationalist community and the wider world, but there are many paths I can imagine it taking. I'm glad that the core concept has resonated with so many people, and looking forward to working together to make it better each year.

Harper's Magazine article on LW/MIRI/CFAR and Ethereum

Cover title: “Power and paranoia in Silicon Valley”; article title: “Come with us if you want to live: Among the apocalyptic libertarians of Silicon Valley” (mirrors: 1, 2, 3), by Sam Frank; *Harper’s Magazine*, January 2015, pg26-36 (~8500 words). The beginning/ending are focused on Ethereum and Vitalik Buterin, so I'll excerpt the LW/MIRI/CFAR-focused middle:

> …Blake Masters-the name was too perfect-had, obviously, dedicated himself to the command of self and universe. He did CrossFit and ate Bulletproof, a tech-world variant of the paleo diet. On his Tumblr’s About page, since rewritten, the anti-belief belief systems multiplied, hyperlinked to Wikipedia pages or to the confoundingly scholastic website Less Wrong: “Libertarian (and not convinced there’s irreconcilable fissure between deontological and consequentialist camps). Aspiring rationalist/Bayesian. Secularist/agnostic/ ignostic . . . Hayekian. As important as what we know is what we don’t. Admittedly eccentric.” Then: “Really, really excited to be in Silicon Valley right now, working on fascinating stuff with an amazing team.” I was startled that all these negative ideologies could be condensed so easily into a positive worldview. …I saw the utopianism latent in capitalism-that, as Bernard Mandeville had it three centuries ago, it is a system that manufactures public benefit from private vice. I started CrossFit and began tinkering with my diet. I browsed venal tech-trade publications, and tried and failed to read Less Wrong, which was written as if for aliens.

> …I left the auditorium of Alice Tully Hall. Bleary beside the silver coffee urn in the nearly empty lobby, I was buttonholed by a man whose name tag read MICHAEL VASSAR, METAMED research. He wore a black-and-white paisley shirt and a jacket that was slightly too big for him. “What did you think of that talk?” he asked, without introducing himself. “Disorganized, wasn’t it?” A theory of everything followed. Heroes like Elon and Peter (did I have to ask? Musk and Thiel). The relative abilities of physicists and biologists, their standard deviations calculated out loud. How exactly Vassar would save the world. His left eyelid twitched, his full face winced with effort as he told me about his “personal war against the universe.” My brain hurt. I backed away and headed home. But Vassar had spoken like no one I had ever met, and after Kurzweil’s keynote the next morning, I sought him out. He continued as if uninterrupted. Among the acolytes of eternal life, Vassar was an eschatologist. “There are all of these different countdowns going on,” he said. “There’s the countdown to the broad postmodern memeplex undermining our civilization and causing everything to break down, there’s the countdown to the broad modernist memeplex destroying our environment or killing everyone in a nuclear war, and there’s the countdown to the modernist civilization learning to critique itself fully and creating an artificial intelligence that it can’t control. There are so many different - on different time-scales - ways in which the self-modifying intelligent processes that we are embedded in undermine themselves. I’m trying to figure out ways of disentangling all of that. . . .I’m not sure that what I’m trying to do is as hard as founding the Roman Empire or the Catholic Church or something. But it’s harder than people’s normal big-picture ambitions, like making a billion dollars.” Vassar was thirty-four, one year older than I was. He had gone to college at seventeen, and had worked as an actuary, as a teacher, in nanotech, and in the Peace Corps. He’d founded a music-licensing

start-up called Sir Groovy. Early in 2012, he had stepped down as president of the Singularity Institute for Artificial Intelligence, now called the Machine Intelligence Research Institute (MIRI), which was created by an autodidact named Eliezer Yudkowsky, who also started Less Wrong. Vassar had left to found MetaMed, a personalized-medicine company, with Jaan Tallinn of Skype and Kazaa, $500,000 from Peter Thiel, and a staff that included young rationalists who had cut their teeth arguing on Yudkowsky's website. The idea behind MetaMed was to apply rationality to medicine-"rationality" here defined as the ability to properly research, weight, and synthesize the flawed medical information that exists in the world. Prices ranged from $25,000 for a literature review to a few hundred thousand for a personalized study. "We can save lots and lots and lots of lives," Vassar said (if mostly moneyed ones at first). "But it's the signal-it's the 'Hey! Reason works!'-that matters. . . . It's not really about medicine." Our whole society was sick - root, branch, and memeplex - and rationality was the only cure. …I asked Vassar about his friend Yudkowsky. "He has worse aesthetics than I do," he replied, "and is actually incomprehensibly smart." We agreed to stay in touch.

One month later, I boarded a plane to San Francisco. I had spent the interim taking a second look at Less Wrong, trying to parse its lore and jargon: "scope insensitivity," "ugh field," "affective death spiral," "typical mind fallacy," "counterfactual mugging," "Roko's basilisk." When I arrived at the MIRI offices in Berkeley, young men were sprawled on beanbags, surrounded by whiteboards half black with equations. I had come costumed in a Fermat's Last Theorem T-shirt, a summary of the proof on the front and a bibliography on the back, printed for the number-theory camp I had attended at fifteen. Yudkowsky arrived late. He led me to an empty office where we sat down in mismatched chairs. He wore glasses, had a short, dark beard, and his heavy body seemed slightly alien to him. I asked what he was working on. "Should I assume that your shirt is an accurate reflection of your abilities," he asked, "and start blabbing math at you?" Eight minutes of probability and game theory followed. Cogitating before me, he kept grimacing as if not quite in control of his face. "In the very long run, obviously, you want to solve all the problems associated with having a stable, self-improving, beneficial-slash-benevolent AI, and then you want to build one." What happens if an artificial intelligence begins improving itself, changing its own source code, until it rapidly becomes - foom! is Yudkowsky's preferred expression - orders of magnitude more intelligent than we are? A canonical thought experiment devised by Oxford philosopher Nick Bostrom in 2003 suggests that even a mundane, industrial sort of AI might kill us. Bostrom posited a "superintelligence whose top goal is the manufacturing of paper-clips." For this AI, known fondly on Less Wrong as Clippy, self-improvement might entail rearranging the atoms in our bodies, and then in the universe - and so we, and everything else, end up as office supplies. Nothing so misanthropic as Skynet is required, only indifference to humanity. What is urgently needed, then, claims Yudkowsky, is an AI that shares our values and goals. This, in turn, requires a cadre of highly rational mathematicians, philosophers, and programmers to solve the problem of "friendly" AI - and, incidentally, the problem of a universal human ethics - before an indifferent, unfriendly AI escapes into the wild.

Among those who study artificial intelligence, there's no consensus on either point: that an intelligence explosion is possible (rather than, for instance, a proliferation of weaker, more limited forms of AI) or that a heroic team of rationalists is the best defense in the event. That MIRI has as much support as it does (in 2012, the institute's annual revenue broke $1 million for the first time) is a testament to Yudkowsky's rhetorical ability as much as to any technical skill.

Over the course of a decade, his writing, along with that of Bostrom and a handful of others, has impressed the dangers of unfriendly AI on a growing number of people in the tech world and beyond. In August, after reading *Superintelligence*, Bostrom's new book, Elon Musk tweeted, "Hope we're not just the biological boot loader for digital superintelligence. Unfortunately, that is increasingly probable." In 2000, when Yudkowsky was twenty, he founded the Singularity Institute with the support of a few people he'd met at the Foresight Institute, a Palo Alto nanotech think tank. He had already written papers on "The Plan to Singularity" and "Coding a Transhuman AI," and posted an autobiography on his website, since removed, called "Eliezer, the Person." It recounted a breakdown of will when he was eleven and a half: "I can't do anything. That's the phrase I used then." He dropped out before high school and taught himself a mess of evolutionary psychology and cognitive science. He began to "neuro-hack" himself, systematizing his introspection to evade his cognitive quirks. Yudkowsky believed he could hasten the singularity by twenty years, creating a superhuman intelligence and saving humankind in the process. He met Thiel at a Foresight Institute dinner in 2005 and invited him to speak at the first annual Singularity Summit. The institute's paid staff grew. In 2006, Yudkowsky began writing a hydra-headed series of blog posts: science-fictionish parables, thought experiments, and explainers encompassing cognitive biases, self-improvement, and many-worlds quantum mechanics that funneled lay readers into his theory of friendly AI. Rationality workshops and Meetups began soon after. In 2009, the blog posts became what he called Sequences on a new website: Less Wrong. The next year, Yudkowsky began publishing *Harry Potter and the Methods of Rationality* at `fanfiction.net`. The Harry Potter category is the site's most popular, with almost 700,000 stories; of these, HPMoR is the most reviewed and the second-most favorited. The last comment that the programmer and activist Aaron Swartz left on Reddit before his suicide in 2013 was on `/r/hpmor`. In Yudkowsky's telling, Harry is not only a magician but also a scientist, and he needs just one school year to accomplish what takes canon-Harry seven. HPMoR is serialized in arcs, like a TV show, and runs to a few thousand pages when printed; the book is still unfinished. Yudkowsky and I were talking about literature, and Swartz, when a college student wandered in. Would Eliezer sign his copy of HPMoR? "But you have to, like, write something," he said. "You have to write, 'I am who I am.' So, 'I am who I am' and then sign it." "Alrighty," Yudkowsky said, signed, continued. "Have you actually read *Methods of Rationality* at all?" he asked me. "I take it not." (I'd been found out.) "I don't know what sort of a deadline you're on, but you might consider taking a look at that." (I had taken a look, and hated the little I'd managed.) "It has a legendary nerd-sniping effect on some people, so be warned. That is, it causes you to read it for sixty hours straight."

The nerd-sniping effect is real enough. Of the 1,636 people who responded to a 2013 survey of Less Wrong's readers, one quarter had found the site thanks to HPMoR, and many more had read the book. Their average age was 27.4, their average IQ 138.2. Men made up 88.8% of respondents; 78.7% were straight, 1.5% transgender, 54.7 % American, 89.3% atheist or agnostic. The catastrophes they thought most likely to wipe out at least 90% of humanity before the year 2100 were, in descending order, pandemic (bioengineered), environmental collapse, unfriendly AI, nuclear war, pandemic (natural), economic/political collapse, asteroid, nanotech/gray goo. Forty-two people, 2.6 %, called themselves futarchists, after an idea from Robin Hanson, an economist and Yudkowsky's former coblogger, for reengineering democracy into a set of prediction markets in which speculators can bet on the best policies. Forty people called themselves reactionaries, a grab bag of former libertarians, ethno-nationalists, Social

Darwinists, scientific racists, patriarchists, pickup artists, and atavistic “traditionalists,” who Internet-argue about antidemocratic futures, plumping variously for fascism or monarchism or corporatism or rule by an all-powerful, gold-seeking alien named Fnargl who will free the markets and stabilize everything else. At the bottom of each year’s list are suggestive statistical irrelevancies: “every optimizing system’s a dictator and i’m not sure which one i want in charge,” “Autocracy (important: myself as autocrat),” “Bayesian (aspiring) Rationalist. Technocratic. Human-centric Extropian Coherent Extrapolated Volition.” “Bayesian” refers to Bayes’s Theorem, a mathematical formula that describes uncertainty in probabilistic terms, telling you how much to update your beliefs when given new information. This is a formalization and calibration of the way we operate naturally, but “Bayesian” has a special status in the rationalist community because it’s the least imperfect way to think. “Extropy,” the antonym of “entropy,” is a decades-old doctrine of continuous human improvement, and “coherent extrapolated volition” is one of Yudkowsky’s pet concepts for friendly artificial intelligence. Rather than our having to solve moral philosophy in order to arrive at a complete human goal structure, C.E.V. would computationally simulate eons of moral progress, like some kind of Whiggish Pangloss machine. As Yudkowsky wrote in 2004, “In poetic terms, our coherent extrapolated volition is our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together.” Yet can even a single human’s volition cohere or compute in this way, let alone humanity’s? We stood up to leave the room. Yudkowsky stopped me and said I might want to turn my recorder on again; he had a final thought. “We’re part of the continuation of the Enlightenment, the Old Enlightenment. This is the New Enlightenment,” he said. “Old project’s finished. We actually have science now, now we have the next part of the Enlightenment project.”

In 2013, the Singularity Institute changed its name to the Machine Intelligence Research Institute. Whereas MIRI aims to ensure human-friendly artificial intelligence, an associated program, the Center for Applied Rationality, helps humans optimize their own minds, in accordance with Bayes’s Theorem. The day after I met Yudkowsky, I returned to Berkeley for one of CFAR’s long-weekend workshops. The color scheme at the Rose Garden Inn was red and green, and everything was brocaded. The attendees were mostly in their twenties: mathematicians, software engineers, quants, a scientist studying soot, employees of Google and Facebook, an eighteen-year-old Thiel Fellow who’d been paid $100,000 to leave Boston College and start a company, professional atheists, a Mormon turned atheist, an atheist turned Catholic, an Objectivist who was photographed at the premiere of *Atlas Shrugged II: The Strike*. There were about three men for every woman. At the Friday-night meet and greet, I talked with Benja, a German who was studying math and behavioral biology at the University of Bristol, whom I had spotted at MIRI the day before. He was in his early thirties and quite tall, with bad posture and a ponytail past his shoulders. He wore socks with sandals, and worried a paper cup as we talked. Benja had felt death was terrible since he was a small child, and wanted his aging parents to sign up for cryonics, if he could figure out how to pay for it on a grad-student stipend. He was unsure about the risks from unfriendly AI - “There is a part of my brain,” he said, “that sort of goes, like, ‘This is crazy talk; that’s not going to happen’” - but the probabilities had persuaded him. He said there was only about a 30% chance that we could make it another century without an intelligence explosion. He was at CFAR to stop procrastinating. Julia Galef, CFAR’s president and cofounder, began a session on Saturday morning with the first of many brain-as-computer metaphors. We are “running rationality on human hardware,” she said, not supercomputers, so the goal was to become incrementally more self-reflective and Bayesian: not perfectly rational agents, but “agent-y.” The workshop’s classes lasted six or so hours a day; activities and conversations went well into the night. We got a condensed treatment of contemporary neuroscience that focused on hacking our brains’ various systems and modules, and attended sessions on habit training, urge propagation, and delegating to future selves. We heard a lot about Daniel Kahneman, the Nobel Prize-winning psychologist whose work on cognitive heuristics and biases demonstrated many of the ways we are irrational. Geoff Anders, the founder of Leverage Research, a “meta-level nonprofit” funded by Thiel, taught a class on goal factoring, a process of introspection that, after many tens of hours, maps out every one of your goals down to root-level motivations-the unchangeable “intrinsic goods,” around which you can rebuild your life. Goal factoring is an application of Connection Theory, Anders’s model of human psychology, which he developed as a Rutgers philosophy student disserting on Descartes, and Connection Theory is just the start of a universal renovation. Leverage Research has a master plan that, in the most recent public version, consists of nearly 300 steps. It begins from first principles and scales up from there: “Initiate a philosophical investigation of philosophical method”; “Discover a sufficiently good philosophical method”; have 2,000-plus “actively and stably benevolent people successfully seek enough power to be able to stably guide the world”; “People achieve their ultimate goals as far as possible without harming others”; “We have an optimal world”; “Done.” On Saturday night, Anders left the Rose Garden Inn early to supervise a polyphasic-sleep experiment that some Leverage staff members were conducting on themselves. It was a schedule called the Everyman 3, which compresses sleep into three twenty-minute REM naps each day and three hours at night for slow-wave. Anders was already polyphasic himself. Operating by the lights of his own best practices, goal-factored, coherent, and connected, he was able to work 105 hours a week on world optimization. For the rest of us, for me, these were distant aspirations. We were nerdy and unperfected. There was intense discussion at every free moment, and a genuine interest in new ideas, if especially in testable, verifiable ones. There was joy in meeting peers after years of isolation. CFAR was also insular, overhygienic, and witheringly focused on productivity. Almost everyone found politics to be tribal and viscerally upsetting. Discussions quickly turned back to philosophy and math. By Monday afternoon, things were wrapping up. Andrew Critch, a CFAR cofounder, gave a final speech in the lounge: “Remember how you got started on this path. Think about what was the time for you when you first asked yourself, ‘How do I work?’ and ‘How do I want to work?’ and ‘What can I do about that?’ . . . Think about how many people throughout history could have had that moment and not been able to do anything about it because they didn’t know the stuff we do now. I find this very upsetting to think about. It could have been really hard. A lot harder.” He was crying. “I kind of want to be grateful that we’re now, and we can share this knowledge and stand on the shoulders of giants like Daniel Kahneman . . . I just want to be grateful for that. . . . And because of those giants, the kinds of conversations we can have here now, with, like, psychology and, like, algorithms in the same paragraph, to me it feels like a new frontier. . . . Be explorers; take advantage of this vast new landscape that’s been opened up to us in this time and this place; and bear the torch of applied rationality like brave explorers. And then, like, keep in touch by email.” The workshop attendees put giant Post-its on the walls expressing the lessons they hoped to take with them. A blue one read RATIONALITY IS SYSTEMATIZED WINNING. Above it, in pink: THERE ARE OTHER PEOPLE WHO THINK LIKE ME. I AM NOT ALONE.

That night, there was a party. Alumni were invited. Networking was encouraged. Post-its proliferated; one, by the beer cooler, read SLIGHTLY ADDICTIVE. SLIGHTLY MIND-ALTERING. Another, a few feet to the right, over a double stack of bound copies of *Harry Potter and the Methods of Rationality*: VERY ADDICTIVE. VERY MIND-ALTERING. I talked to one of my roommates, a Google scientist who worked on neural nets. The CFAR workshop was just a whim to him, a tourist weekend. “They’re the nicest people you’d ever meet,” he said, but then he qualified the compliment. “Look around. If they were effective, rational people, would they be here? Something a little weird, no?” I walked outside for air. Michael Vassar, in a clinging red sweater, was talking to an actuary from Florida. They discussed timeless decision theory (approximately: intelligent agents should make decisions on the basis of the futures, or possible worlds, that they predict their decisions will create) and the simulation argument (essentially: we’re living in one), which Vassar traced to Schopenhauer. He recited lines from Kipling’s “If-” in no particular order and advised the actuary on how to change his life: Become a pro poker player with the $100k he had in the bank, then hit the Magic: The Gathering pro circuit; make more money; develop more rationality skills; launch the first Costco in Northern Europe. I asked Vassar what was happening at MetaMed. He told me that he was raising money, and was in discussions with a big HMO. He wanted to show up Peter Thiel for not investing more than $500,000. “I’m basically hoping that I can run the largest convertible-debt offering in the history of finance, and I think it’s kind of reasonable,” he said. “I like Peter. I just would like him to notice that he made a mistake . . . I imagine a hundred million or a billion will cause him to notice . . . I’d like to have a pi-billion-dollar valuation.” I wondered whether Vassar was drunk. He was about to drive one of his coworkers, a young woman named Alyssa, home, and he asked whether I would join them. I sat silently in the back of his musty BMW as they talked about potential investors and hires. Vassar almost ran a red light. After Alyssa got out, I rode shotgun, and we headed back to the hotel.

It was getting late. I asked him about the rationalist community. Were they really going to save the world? From what? “Imagine there is a set of skills,” he said. “There is a myth that they are possessed by the whole population, and there is a cynical myth that they’re possessed by 10% of the population. They’ve actually been wiped out in all but about one person in three thousand.” It is important, Vassar said, that his people, “the fragments of the world,” lead the way during “the fairly predictable, fairly total cultural transition that will predictably take place between 2020 and 2035 or so.” We pulled up outside the Rose Garden Inn. He continued: “You have these weird phenomena like Occupy where people are protesting with no goals, no theory of how the world is, around which they can structure a protest. Basically this incredibly, weirdly, thoroughly disempowered group of people will have to inherit the power of the world anyway, because sooner or later everyone older is going to be too old and too technologically obsolete and too bankrupt. The old institutions may largely break down or they may be handed over, but either way they can’t just freeze. These people are going to be in charge, and it would be helpful if they, as they come into their own, crystallize an identity that contains certain cultural strengths like argument and reason.” I didn’t argue with him, except to press, gently, on his particular form of elitism. His rationalism seemed so limited to me, so incomplete. “It is unfortunate,” he said, “that we are in a situation where our cultural heritage is possessed only by people who are extremely unappealing to most of the population.” That hadn’t been what I’d meant. I had meant rationalism as itself a failure of the imagination. “The current ecosystem is so totally fucked up,” Vassar said. “But if you have conversations here”-he gestured at the hotel-“people

> change their mind and learn and update and change their behaviors in response to the things they say and learn. That never happens anywhere else.” In a hallway of the Rose Garden Inn, a former high-frequency trader started arguing with Vassar and Anna Salamon, CFAR’s executive director, about whether people optimize for hedons or utilons or neither, about mountain climbers and other high-end masochists, about whether world happiness is currently net positive or negative, increasing or decreasing. Vassar was eating and drinking everything within reach. My recording ends with someone saying, “I just heard ‘hedons’ and then was going to ask whether anyone wants to get high,” and Vassar replying, “Ah, that’s a good point.” Other voices: “When in California . . .” “We are in California, yes.”
>
> …Back on the East Coast, summer turned into fall, and I took another shot at reading Yudkowsky’s Harry Potter fanfic. It’s not what I would call a novel, exactly, rather an unending, self-satisfied parable about rationality and trans-humanism, with jokes.
>
> …I flew back to San Francisco, and my friend Courtney and I drove to a cul-de-sac in Atherton, at the end of which sat the promised mansion. It had been repurposed as cohousing for children who were trying to build the future: start-up founders, singularitarians, a teenage venture capitalist. The woman who coined the term “open source” was there, along with a Less Wronger and Thiel Capital employee who had renamed himself Eden. The Day of the Idealist was a day for self-actualization and networking, like the CFAR workshop without the rigor. We were to set “mega goals” and pick a “core good” to build on in the coming year. Everyone was a capitalist; everyone was postpolitical. I squabbled with a young man in a Tesla jacket about anti-Google activism. No one has a right to housing, he said; programmers are the people who matter; the protesters’ antagonistic tactics had totally discredited them.
>
> …Thiel and Vassar and Yudkowsky, for all their far-out rhetoric, take it on faith that corporate capitalism, unchecked just a little longer, will bring about this era of widespread abundance. Progress, Thiel thinks, is threatened mostly by the political power of what he calls the “unthinking demos.”

Pointer thanks to /u/Vulture.

Moving towards the goal

This is a linkpost for http://mindingourway.com/moving-towards-the-goal/

This post contains some advice. I dare not call it obvious, as the illusion of transparency is ever-present. I will call it simple, but people occasionally remind me that they really appreciate the simple advice. So here we go:

1

(As usual, this advice is not for everyone; today I am primarily speaking to those who have something to protect.)

I have been spending quite a bit of time, recently, working with people who are explicitly trying to hop on a higher growth curve and have a larger impact on the world. (Most of them effective altruists.) They wonder how the big problems can be solved, or how one single person can themselves move the needle in a meaningful way. They ask questions like "what needs to be done?", or "what sort of high impact things can I do right now?"

I think this is the wrong way of looking at things.

When I have a big problem that I want solved, I have found that there is one simple process which tends to work. It goes like this:

1. Move towards the goal.

(It's simple, not easy.)

If you follow this process, you either win or you die. (Or you run out of time. Speed is encouraged. So are shortcuts, so is cheating.)

The difficult part is hidden within step 1: it's often hard to keep moving towards the goal. It's difficult to stay motivated. It's difficult to stay focused, especially when pursuing an ambitious goal such as "end ageing," which requires overcoming some fairly significant obstacles.

But we are human beings. We are the single most powerful optimization process in the known universe, with the only exception being *groups* of human beings. If we set ourselves to something and don't stop, we either suceed or we die. There's a whole slew of advice which helps make the former outcome more likely than the latter (via efficiency, etc.), but first it is necessary to begin.

Moving towards the goal doesn't mean you have to work directly on whatever problem you're solving. If you're trying to end aging, then putting on a lab coat and combining random chemicals likely won't do you much good.

Rather, moving towards the goal is about *always acting to solve the problem,* with each motion. Identify the path to the goal that seems shortest, and then walk it. Maybe you need to acquire financial stability first, and more knowledge second. Maybe you need to expand your social network, or fulfill your social attachment needs. Maybe you need to acquire a new skill. Maybe you have *no idea* how to start, in which case you need to gain more information, do some thinking, and gain a higher vantage point from which to search for a path to the goal.

But no matter what, there is always *some* way to keep moving towards the goal. Get stronger. Get smarter. Return with allies at your back.

2

Here's the pattern that this advice is designed to work against: consider the effective altruist, asking "what needs to be done?", or "what sort of high impact things can I do right now?"

I expect people to go much farther by first identifying an actual goal, and then moving towards it. Which breaks my one-step advice above into a more practical two-step process:

Step 1: identify the goal. Figure out what you're actually trying to accomplish. Probe your motivations, and trace them back to something that compels.

I'm not suggesting tracing your motivations all the way up to "final" goals; it's a bit presumptuous to claim knowledge of "final goals" given modern introspective capabilities. Rather, look for important problems that you're trying to solve in the world today.

For example, you might be trying to fix education, end hunger, eliminate a disease, prevent aging, become immortal, end suffering, prevent human extinction, or whatever. None of these are *ends unto themselves*, but they're all problems that need solving.

Identifying a goal that compels—that really needs to be solved, and that won't be solved (or won't be solved fast enough) by default—is not always an easy task. Many people are locked into a mindset where they couldn't possibly actually solve any big problems, because big problems are big and people are small. Breaking out of that mindset is a topic for another day; for now I'll assume you have picked your poison and identified some goal to achieve, even if only a minor one.

Step 2: move towards it. So, you've found a goal. Nice work.

Now solve it tomorrow.

Can you? Seriously ask yourself whether or not you can solve the problem tomorrow. I don't care how ambitious it is. Can you solve it tomorrow? If yes, then do it. If not, why not? Say the obstacles aloud.

The usual answers are something like "I lack the power, time, money, network, and so on." Which is great! Now we're getting somewhere.

These are what you need to work on tomorrow, if you want to solve the problem.

Don't ask "what would be good to do," ask "what is standing between me and solving the problem immediately." Identify the obstacles. Your task is now to either remove them or cheat your way around them.

Of course, most of the obstacles themselves are still too big and vague. So ask yourself why you can't solve those problems tomorrow. Say you don't know the people you'd need to know to have a shot at fixing education. Can you contact them all tomorrow? That *probably* wouldn't go well, but why not? What are the obstacles between you and acquiring the resources you're going to need?

Rinse, repeat. Identify the obstacles to overcoming the obstacles, and so on. Eventually, this process will ground out in things that you can actually start doing tomorrow, with a path that you can trace all the way back up to your goal.

Once you have that, throw reservations to the wind, and start *today.*

3

Moving towards the goal doesn't solve the whole problem. If you want to solve a goal effectively, in the time allotted, it is important to approach the obstacles in the right order, to identify the ones you can safely cheat past, to correctly distinguish between short paths to the goal and long ones. But many people aren't there yet: they're still asking "what would be good for me to do," and not "what stands between me and solving the whole problem tomorrow."

My advice, if you want to be effective, is *always be solving the problem.* With each motion, be overcoming an obstacle that stands between you and the goal. If the obstacles are too large, then your next task is to get stronger, get smarter, or find a way around. That is what it means, to find a path to the goal.

To achieve a goal, simply keep moving along that path.

MIRI's technical research agenda

I'm pleased to announce the release of Aligning Superintelligence with Human Interests: A Technical Research Agenda written by Benja and I (with help and input from many, many others). This document summarizes and motivates MIRI's current technical research agenda.

I'm happy to answer questions about this document, but expect slow response times, as I'm travelling for the holidays. The introduction of the paper is included below. (See the paper for references.)

The characteristic that has enabled humanity to shape the world is not strength, not speed, but intelligence. Barring catastrophe, it seems clear that progress in AI will one day lead to the creation of agents meeting or exceeding human-level general intelligence, and this will likely lead to the eventual development of systems which are "superintelligent'' in the sense of being "smarter than the best human brains in practically every field" (Bostrom 2014). A superintelligent system could have an enormous impact upon humanity: just as human intelligence has allowed the development of tools and strategies that let humans control the environment to an unprecedented degree, a superintelligent system would likely be capable of developing tools and strategies that give it extraordinary power (Muehlhauser and Salamon 2012). In light of this potential, it is essential to use caution when developing artificially intelligent systems capable of attaining or creating superintelligence.

There is no reason to expect artificial agents to be driven by human motivations such as lust for power, but almost all goals can be better met with more resources (Omohundro 2008). This suggests that, by default, superintelligent agents would have incentives to acquire resources currently being used by humanity. (Can't we share? Likely not: there is no reason to expect artificial agents to be driven by human motivations such as fairness, compassion, or conservatism.) Thus, most goals would put the agent at odds with human interests, giving it incentives to deceive or manipulate its human operators and resist interventions designed to change or debug its behavior (Bostrom 2014, chap. 8).

Care must be taken to avoid constructing systems that exhibit this default behavior. In order to ensure that the development of smarter-than-human intelligence has a positive impact on humanity, we must meet three formidable challenges: How can we create an agent that will reliably pursue the goals it is given? How can we formally specify beneficial goals? And how can we ensure that this agent will assist and cooperate with its programmers as they improve its design, given that mistakes in the initial version are inevitable?

This agenda discusses technical research that is tractable today, which the authors think will make it easier to confront these three challenges in the future. Sections 2 through 4 motivate and discuss six research topics that we think are relevant to these challenges. Section 5 discusses our reasons for selecting these six areas in particular.

We call a smarter-than-human system that reliably pursues beneficial goals "aligned with human interests" or simply "aligned." To become confident that an agent is aligned in this way, a practical implementation that merely *seems* to meet the challenges outlined above will not suffice. It is also necessary to gain a solid theoretical understanding of why that confidence is justified. This technical agenda argues that there is foundational research approachable today that will make it easier to develop aligned systems in the future, and describes ongoing work on some of these problems.

Of the three challenges, the one giving rise to the largest number of currently tractable research questions is the challenge of finding an agent architecture that will reliably pursue the goals it is given—that is, an architecture which is alignable in the first place. This requires theoretical knowledge of how to design agents which reason well and behave as intended even in situations never envisioned by the programmers. The problem of highly reliable agent designs is discussed in Section 2.

The challenge of developing agent designs which are tolerant of human error has also yielded a number of tractable problems. We argue that smarter-than-human systems would by default have incentives to manipulate and deceive the human operators. Therefore, special care must be taken to develop agent architectures which avert these incentives and are otherwise tolerant of programmer error. This problem and some related open questions are discussed in Section 3.

Reliable, error-tolerant agent designs are only beneficial if they are aligned with human interests. The difficulty of concretely specifying what is meant by "beneficial behavior" implies a need for some way to construct agents that reliably *learn* what to value (Bostrom 2014, chap. 12). A solution to this "value learning'' problem is vital; attempts to start making progress are reviewed in Section 4.

Why these problems? Why now? Section 5 answers these questions and others. In short, the authors believe that there is theoretical research which can be done today that will make it easier to design aligned smarter-than-human systems in the future.

Kickstarting the audio version of the upcoming book "The Sequences"

LessWrong is getting ready to release an actual book that covers most of the material found in the Sequences.

There have been a few posts about it in the past, here are two: the title debate, content optimization.

We've been asked if we'd like to produce the audiobook version and the answer is yes. This is a large undertaking. The finished product will probably be over 35 hours of audio.

To help mitigate our risk we've decided to Kickstarter the audiobook. This basically allows us to pre-sell it so we're not stuck with a large production cost and no revenue.

The kickstarter campaign is here: https://www.kickstarter.com/projects/1267969302/lesswrong-the-sequences-audiobook

If you haven't heard of us before we've already produced some sequences into audiobooks. You can see them and listen to samples which are indicative of the audio quality here.