Best of LessWrong: December 2015

Best of LessWrong: December 2015

Why CFAR? The view from 2015

Follow-up to: 2013 and 2014.

In this post, we:

- Revisit CFAR's mission, and why that mission matters today;
- Review our progress to date;
- Offer a look at our financial overview;
- Share our ambitions for 2016; and
- Ask your help, via donations and other means.

We are in the middle of our matching fundraiser; so if you've been considering donating to CFAR this year, now is an unusually good time.

CFAR's mission, and why that mission matters today

CFAR's mission is to help people develop the abilities that let them meaningfully assist with the world's most important problems, by improving their ability to arrive at accurate beliefs, act effectively in the real world, and sustainably care about that world.

We know this is an audacious thing to try—especially the "ability to form accurate beliefs" part—but it seems to us that such attempts work sometimes anyhow. Eliezer's Sequences seem to offer principled improvements to some aspects of some peoples' world-modeling skill (via synthesizing much recent cognitive science, probability theory, etc.); this seems to us to be a useful point from which to build.

The fact remains that we do not yet have the talent necessary to win—to see the world's problems clearly, plot strategies that have a shot at working, update when those strategies don't work, and plan effectively around unknowns. To avoid any great filters that may be lurking, solve global and even astronomical challenges, and create a flourishing world for all.

Arguably, people of the caliber we're shooting for don't exist yet, but even if they do, it seems clear that we don't have enough of them to have enough of a guarantee of *actually succeeding*.

So, audacious or not, this is a task that needs to be done, and CFAR is our attempt to do it. If we can widen the bottleneck on thinking better and doing more, we're increasing the odds of a better future regardless of what the important problems turn out to be.

Our progress to date

By the end of 2014, CFAR had created workshops that participants liked a lot and which evidence suggests had concrete benefits for them. However, our mission remains to impact the world. The question became whether we could adapt our workshops into something that had the potential for large impact.

Our central goal for 2015 was therefore to create what we called a "minimum *strategic* product" -- a product that, as we put it last year, would "more directly justify

CFAR's claim to be an effective altruist project" by demonstrating that we could sometimes improve peoples' thinking skill, competence, and/or do-gooding to the point where they were able to engage in direct work on a key talent-limited task.

Running the MIRI Summer Fellows program gave the opportunity we'd sought to try our hand at creating such direct impact.   Our plan was to test and develop our curriculum and training methods through running a training program that would not only improve people’s ability to think about some of the big questions, but also do so in a fashion that could lead to immediate progress.

How did we do? Here’s what Nate Soares, MIRI’s Executive Director, had to say:

> “MSFP was a resounding success: many participants gained new skills relevant to alignment research, and the program led directly to multiple MIRI hires. The world needs more talented people focusing on big important problems, and CFAR has figured out how to develop those sorts of talents in practice.”

While working to help create AI alignment researchers, we also found that this focus on how to become a better scientist led us into more fruitful territory for improving our understanding of the art. (If you're curious, you can see a highly incoherent version of some of the skills we tried to get across in this working document. Read below for more details about art creation, and our plans to expand on more targeted training programs.)

Last year's "goals for 2015"

We hit some of our concrete goals for 2015 and got distracted from others (partly, perils of unanticipated opportunities :-/).

We created a provisional metric for participants' before-and-after strategic usefulness, hitting the first goal; we started tracking that metric, hitting the second goal. Then we found that the metric was too unwieldy and too interpersonally tricky to regularly use on participants, making this "hitting" of our "goals" somewhat less useful than we had hoped. (On the upside, we learned something about how not to build metrics. :-/)

We then got the opportunity to run MIRI Summer Fellows, as noted above... and mostly dropped our previously declared goals to pull off the program, partly because our goals had been meant as a concretization of "can we train people who matter for the world", and the Summer Fellows program seemed like a better concretization of the same. (The program required a *lot* of new curriculum beyond what we had before, and a lot of skill development on the part of our teaching staff; and even so, and despite Nate's calling it a "resounding success", we had a feeling of leaving a lot of opportunity on the table; opportunity we intend to pick up in our second MIRI Summer Fellows program this coming summer).

From the original "concrete goals" list: goal three was a bit wishy-washy but was probably done. Goals four and five we did not even measure to see if we hit. We should and will measure this, and will let you know when we do; it seems good that we opportunistically put our all into the summer fellows program (and okay to de-emphasize old goals in pursuit of that), but good also to then follow it up for the sake of feedback loops and honesty.

Organizational capital

2015 was the year in which we finally managed to stop wearing all the hats thanks to a huge increase in organizational capital. At the start of 2015 workshops were stressful for staff. Between workshops, our workdays were cluttered with a disproportionate amount of attention spent on logistics, alumni followups, and tasks like accounting.

This stress and clutter was part of what was preventing us from seeing what we were doing, and figuring out how to actually contribute to the world; smoothing out the wrinkles in our day-to-day workflow was (we think) a major stepping stone toward discovering our minimum strategic product.

That's why we spent a lot of time and effort this year on streamlining operations and increasing specialization so that we could both free the capacity to focus on developing the art and create the capacity to scale our workshops. We systematized tasks like accounting and venue searches, and began using alumni volunteers as follow up mentors to supplement our newly-created post-workshop email exercises and online hangouts. These efforts culminated in two new hires—Pete Michaud and Duncan Sabien—and a reorganization of CFAR into two subteams, Core (focused on operations) and Labs (focused on research).

For a complete overview of what we intend to accomplish in 2016, see Ambitions for 2016 below.

Some snapshots from our rationality development

There is the process by which we improve a workshop, and there is the process by which we improve our understanding of how rationality works at its core. The two processes don't always help one another, but this year they did.

How we got there:

- As it turns out, attempting to create AI risk scientists (as opposed to boosting the scientist-nature of everyday people) put a subtle but very different spin on the teaching of Sequences-style epistemic rationality. It helped that the researchers were themselves trying to model mind-like processes and that they stubbornly insisted on building related models of what the heck we were trying to convey.
- MIRI Summer Fellows was also a project we could just actually see mattered, and there's nothing quite like actual stakes when it comes to creating a sense of drive and purpose, and being willing to update.
- Improving organizational capital created a positive feedback loop. Working to make our workshops "crisp"—to clean up the methods and metaphors that weren't pulling their weight—helped make more of what we knew more visible.

Here are some brief highlights of the new *Art of Rationality* that we're currently seeing:

- **One pillar, not three.** CFAR has long talked about wanting to boost three distinct things in our participants (competence, epistemic rationality, and do-gooding). But we've had the strong sense that there were ways to strengthen all three through the practice of a single, unified art of "applied rationality" (for instance, a deep understanding of reductionism seems to help with all three). Recently, we've gotten better at articulating *how* this link works. For example:

- **Double Crux** is a structured format for collaboratively finding the truth in cases where two people disagree. Instead of non-interactively offering pieces of their respective platforms, people jointly seek the actual question at the crux of the disagreement—the root uncertainty that has the potential to affect *both* of their beliefs. We introduced this as an epistemic rationality technique, and used in in this way at e.g. EA Global, where people argued about cause prioritization; it then made its way also into our material on competence and on how to sustainably care deeply about the world. (See the next two bullet points.)
- **Competence *as* "deep/internal epistemic rationality."** If I am frequently late to appointments and "don't want to be," one can frame this as stemming from an inaccurate anticipation somewhere in my mind—perhaps I mis-anticipate whether my actions will make me late, or perhaps I disagree with myself as to whether lateness in fact harms my goals. Either way, it can be helpful (in our experience) to "internally double crux" the apparent disagreement (i.e., to play the double crux game between two different models within my own head, working until I have both a better model and a better actual outcome). More generally, we are increasingly making headway on "competence" or "instrumental rationality" problems via techniques aimed at integrating accurate beliefs into all parts of one's psyche.
- **Do-gooding and epistemic rationality.** "Do-gooding" would seem to be a goal that some have and others don't, and it would seem odd to try to shift *goals* by learning epistemic rationality. But it seems to many of us (informally, anecdotally) that there is a kind of "deep epistemic rationality" that doesn't *change* one's goals, but *does* help one make actual contact with what is at stake in the world, and with the parts of one's psyche that *already* care about those stakes... and this can sometimes help in practice to build deep, sustainable caring. The idea is again to e.g. notice a part of you that thinks the world matters, and a part of you that is afraid to look in that direction, and help these parts trade model-pieces and update back and forth (double crux, again). For an early attempt to articulate pieces of this "art of connecting to deep caring", see Val's recent post on grieving.
- **Teaching the synthesis.** Our pre-2015 workshops were made of techniques, which was like sounding out words a letter at a time (C-A-T…C…Ca…Cat!). After years of trying to use these techniques to point at the deeper skill (Cat! Hat! Antidisestablishmentarianism!), we've finally found framings and explanations (like this one) that actually bridge the gap. Those framings, plus an explicit emphasis on synthesis and the addition of peer-to-peer tutoring, have successfully transformed the techniques into stepping stones toward the actual art. (The techniques are now stuffed into the first two days; the synthesis, and the rhythms of using applied rationality in practice, now occupy the second half of the workshop and give people a better sense of the lived feeling of the art. We think.)

This is the beginning of work that we're poised to expand and improve in the coming year via our new Labs group.

Financial Retrospective for 2015

General overview

Our net cashflow for the year is about $14k positive so far, though without any further revenue we expect to be around $30k negative by the end of December 2015, as most of our large expenses (rent, payroll, etc.) occur at the end of the month. Note that this includes donation revenue from last year's winter fundraiser.

Our basic monthly operating costs for 2015 have averaged $40k, although the average after September went up to $44k due to changing and slightly expanding our team. This is the number we use to determine burn rate.

$30k of this was payroll in the last quarter, and the rest was split amongst rent and utilities, parking, office supplies, meals, and miscellaneous. Many of these resources are used for in-office events like test sessions, Less Wrong meetups, and rationality training sessions; each staff member has a different and often changing split of percentage time working on operations, curriculum design, teaching, data analysis, etc. That's why giving a good number for monthly overhead is tricky and unreliable. But to give it a go, it looks like roughly a third of monthly expenses is for organization maintenance.

A bit over half of the revenue covering this came from donations. The rest came from net revenue from our standard introductory workshops plus MIRI's payment for our running MSFP. (More details below.)

Main workshops

Our standard introductory workshops serve several important purposes for us. One of them is that we hope to develop useful products that simultaneously support our mission and also make CFAR less fiscally dependent on donations.

We ran four of these workshops (three in the Bay Area and one in Boston). They varied widely in both cost and revenue due to travel, testing out new venues, changing the number of participants per workshop, and several other factors. All told, ignoring costs of staff time (as that's factored into the above burn rate), CFAR main workshops took in a total of ~$123k *net* revenue (i.e., revenue exceeding cost), or an average of ~$31k net revenue per workshop. Compared to last year, this is down ~$107k total, but up ~$6k per workshop. This is due to us choosing to run less than half as many workshops so as to focus on:

1. Making the workshops more efficient
2. Running other programs equally well
3. Setting up better systems both for workshops and for research

In addition, we've continued a trend from last year: we've decreased the per-workshop cost in staff time, partly through streamlined curriculum and improved systems and partly through training volunteers to conduct follow-ups, freeing up our core staff to build new programs and spend more time developing advanced rationality theory and instruction. (The volunteer training also does double-duty: the original impetus for doing it was wanting to help alumni benefit from the "learn by teaching" phenomenon, so we are both freeing up staff time and also using this to help deepen alums' skill with rationality.)

Alumni events

CFAR typically goes into alumni events (workshops and the annual reunion) with the assumption that we’re taking on a cost. We view these as opportunities to explore potentially new areas of rationality and also as ways of encouraging and supporting the CFAR alumni community in their development as rationalists and as a community. It has generally been our policy that we don’t charge for alumni events, but instead we let our alumni know what the per capita cost comes to and ask them to consider donating to compensate.

We track the donations that are in support of these events separately from our standard general donations. As a result, we can pretty clearly see how much each event cost us on beyond the associated donations. That is, we can see net cost. In that spirit, here is what we “paid” on net for each of our alumni programs, ignoring staff time:

- For net zero cost (participants covered meals), we ran a one-day workshop out of the CFAR office on applying Sequences-style thinking to one’s daily life and to hard problems like exrisk, as part of our preparation for MSFP.
- For net zero cost (again, participants covered meals), we ran a 2-day workshop out of the CFAR office on applying Sequences-style thinking to AI risk analysis, also as part of our preparation for MSFP.
- For net zero cost (participants donated enough to cover venue and meals), we ran a “Hamming” workshop in Boston, to explore what techniques are needed to identify and dive into the most important problems one is currently facing (at work, in one’s personal life, as an altruist, or in whatever other domain).
- For ~$2k, we ran a mentoring workshop out of Tiburon, to train volunteers to help us run large-scale workshops and also to do follow-up conversations with participants to help them benefit from the workshop in the weeks & months afterwards.
- For ~$15k, we ran our annual alumni reunion. This year we had ~130 participants, with presentations and exercises on some angles on rationality that we think are promising. These also seem to be a lot of fun and help to energize the alumni community and keep us in touch with fresh ideas from the community that haven’t yet been put in writing.
- For net zero cost, we have continued to run a weekly “rationality dojo” out of the CFAR office, where alumni work to deepen their skills with rationality and experiment with possible refinements or additions to the art.

Special programs

This year we ran two main summer programs:

- SPARC ran for its fourth year in a row. Cisco and MIRI covered the costs of this program, so the non-time cost to CFAR was nil.
- MIRI hired CFAR to run a three-week intensive Summer Fellows Program (MSFP), aimed at identifying and developing promising math research talent potentially related to AI safety research. MIRI covered the costs of running MSFP and paid CFAR $85k to cover both curriculum development time and time running the program itself.

In addition, an unnamed company hired CFAR to run a small training for them. The net financial effect on CFAR was zero: we charged enough to cover costs, viewing this workshop as an opportunity to continue exploring how CFAR might tailor its material for particular workplaces or specific needs.

Financial Summary

Our financial focus this last year was less on making money *now* and more on establishing internal infrastructure and strategies for developing solid income *going forward*.

We're now in an excellent position to make CFAR much less dependent on donations going forward while simultaneously putting more focused effort on development, testing, and sharing of rationality tools than we've been able to do in the past.

This has made 2016 look very promising — but it has also put us in a difficult position right now.

We're farther behind right now than we were this time last year, and we need some capital to implement the plans we have in mind. Predicting markets is always hard, but we think that with one more financial push this winter, we can both improve our contribution to the development of rationality and also make CFAR largely or maybe even entirely financially self-sustaining in 2016.

Ambitions for 2016

Hitting Scale

CFAR's mission cashes out when people we equipped to think better and do more are actually in positions where they are changing the future of our world for the better.

With our external brand and our positioning within the community, we are perhaps uniquely well positioned to attract bright people, orient them to the values of systematically truer beliefs and world scale impact, and then make sure they get into the highest leverage positions they can fill.

We've spent the last three years leveling up our own ability to transmit a skillset and culture that we believe will move the needle in the right direction, and now is the time to execute at scale.

Core and Labs

To make scaling possible and still be able to competently tackle the pedagogical challenges we face, CFAR has arranged itself into two divisions: CFAR Core and CFAR Labs.

Pete Michaud (that's me!) was hired to manage Core operations, including workshop and curriculum production and logistics. Anna Salamon will take the helm of CFAR Labs, which will be principally responsible for answering the questions:

- What are the highest impact skillsets?
- How can we detect them?
- How can we train them?
- Is our training actually affecting the important dimensions at the high end?

The Plan

Broadly, in order to attract more people, level them up reliably, and make sure they land in the highest impact positions they can, our plan is to:

1. Substantially increase workshop volume
2. Expand our community and continued training opportunities
3. Directly address talent gaps by working with other organizations
4. Continue increasing the quality of our instruction

Increase Workshop Volume

We intend to substantially increase the number of intake workshops we run *and* the number of participants we can serve per workshop.

“Intake workshops” here means workshops for people who haven’t necessarily been exposed to our material or community; said another way, these are workshops that will bring new people into our alumni network.

We are actively seeking a direct sales manager who can not only generate leads but close workshop sales. An alternative is to hire a two person marketing and sales team who together can generate leads and place prospects into workshops.

With the help of that new outreach team, we hope to add on the order of 1,000 new alumni in 2016, increasing our total throughput by nearly an order of magnitude.

Handling that new volume of alumni will require increasing attention to streamlining operations, which CFAR Core is handling partially by adding new team members and clarifying roles. In addition to me as the new Managing Director, we’ve already hired Duncan Sabien, an experienced educator and robustly capable operations generalist. Aside from the outreach team already mentioned, we also intend to hire a community manager (see below for details) and office assistant to fill in the inevitable gaps of an organization moving as fast as we intend to.

Community and Continued Training Opportunities

Bringing more talented people into the alumni network is only half the battle. Once participants have gone from “Zero to One,” only a community of practice can help ensure continued growth for most people.

We believe that one of the primary benefits of CFAR training is ongoing participation in the alumni community, both local to the Bay Area and throughout the world in local meetups and online. That’s why we’re going to invest in making the community stronger, with even more alumni events, experimental workshops, and deep-dive classes into specific aspects of our curriculum.

Perhaps the crown jewel of our community program is our Mentorship Training Program (MTP), which began its life as our TA Workshop. We intend to develop that seed into a robust pipeline capable of transforming workshop participants into trained rationality instructors.

One major benefit of the MTP will be that we’ll have more mentors and instructors to handle the increased load of all these workshops, classes, and other events.

But the MTP is a major growth opportunity even for people who aren’t necessarily interested in spreading the art of rationality themselves; we believe from our experience over the past three years that the best way to fully grok the art is to be immersed in a field of peers striving for the same, and ultimately to be able to teach it yourself.

This is what we intend to create with the MTP and new focus on community events.

To plan and manage all these alumni events, we're looking for a capable community manager.

Directly Addressing Talent Gaps

In addition to our classic workshops and general education alumni programs, we'll also be attempting to ramp up our targeted workshops meant to fill talent gaps for specific organizations.

For example, we'll run our second MIRI Summer Fellows Program, as well as a grant funded by the Future of Life Institute to help promising upcoming AI researchers think about AI safety. We're in conversation with other organizations, and it's our intention to have an increasing number of these workshops that focus on thinking skills needed for particular tasks in order to help fill critical gaps in important organizations on very small time horizons.

If funding permits and our experiments in this area go well, we intend to make these types of workshops more frequent, and perhaps expand on past success with programs like a European SPARC, and possible "summer camp" style events where we try to identify particularly talented high school students for training and recruitment into existential risk research.

Labs: Informal experimentation toward a better "Applied Rationality"

The split between Core and Labs doesn't only allow focus on operations--it also allows our Lab folk to invest in the informal experiments, arguments, data-gathering, etc. that seems, over time, to conduce to a better applied rationality.

(This process is messy. Rationality today is not at the level of Newton. It isn't even at the level of Ptolemy, who, despite the mockability of the nested-epicycles method, could predict the motions of the planets with great precision. Rationality is more at the level of a toddler running around, putting everything in its mouth, and ending up thereby with a more integrated informal world-model by having examined many example-objects through several senses each. Our aim this year in Labs is basically to put many many things in our mouths rapidly, and to argue about models in between, and to especially expose ourselves to people who are working on issues that matter in already-very-competent ways who we can nevertheless try to make better, and to try in this way to get a better sense of the higher-end parts of "rationality".)

Toward this end, Labs is currently:

- Offering one-on-one coaching to quite a few individuals who seem to be contributing to the world in a high-end way; and trying to figure out how they're doing what they're doing, and what pieces may help them contribute more;
- Working toward more robust and explicit models of the underlying mechanisms that create drive, scientific and epistemic skill, and relevant real-world competence (and how to intervene upon them);

- Creating new written rationality sequences meant to expand upon, augment, and improve the original sequences that brought so many people into the culture of being “less wrong,” and oriented them around audacious goals that actually make a difference;
- Planning experimental workshops of varied sorts, aiming to boost people further toward "actually useful skill-levels in applied rationality".

We are very excited, and expect that art development will be much easier now that we have a subteam who is free to just actually focus on it. (Last year, we were all doing workshop admissions, logistics, accounting, ...)

Limitations and Updates

The primary limiting factor in these plans is our ability to attract a truly excellent sales person or team. With a sufficient workshop participation, cashflow bottlenecks are broken and we‘ll achieve economies of scale that will fundamentally transform our operations.

Failing that recruitment, the next best alternative is to grow organically through the MTP and other community programs. That is a much slower process, but pushes us in the same fundamental direction.

And as always, our plans coming into contact with the reality of 2016 will correctly cause us to update, iterate, and potentially pivot given new evidence and insight.

The path forward, and how you can help

CFAR’s mission is to gather together people with the potential for real and meaningful impact, and to cause them to come closer to meeting that potential. It doesn’t much matter whether you think we’re under a ticking clock of existential risk, or you’re concerned about a million humans dying every week, or you’re simply grumpy that we haven’t gotten a human past low earth orbit since 1972—our individual and collective thinking skill is a key bottleneck on our future.

Applied rationality, more than almost anything else, has a shot at being a *truly* all-purpose tool in humanity’s toolkit, and the bigger the problems on the horizon, the more vital that tool becomes.

2016 will be a particularly critical year in CFAR’s history. We’re restructuring our team in pretty major ways, and finding the right team members (or not) will determine our ability to get the right character and culture from this new beginning; and we've had at least three good people in the last eight months who we wanted to hire, and who wanted to work for us, but who required salaries we couldn't afford. Beginnings are far easier times in which to make change, and this is the closest we've come to a fresh beginning -- and the time we've most expected differential impact from marginal donation -- since our inaugural fundraiser of late 2012.

The world of AI risk is changing rapidly, and decisions made over the coming months will shape the future of the field -- it would be well to get relevant training programs going *now,* and not to wait for some later additional hard-won new beginning for CFAR in 2018 or something. The strategic competence we will have going into the spring is likely to be the difference between a CFAR that actually matters, and one that sounds good but is ultimately irrelevant.

There are at least four major ways to help:

1. Donate directly to our winter fundraising drive. This is the most straightforward way to help, and makes a categorical difference in our ability to execute the mission. (A large majority of our funding comes from small donors.)
2. If you're interested in rationality, or in the larger questions of humanity's future and existential risk, consider reading the Sequences, or otherwise working to improve your thinking and world-modeling skill. (Strong community epistemology is extremely helpful.)
3. We're always looking for new alumni, particularly those who care about both rationality and the world. If you haven't been, consider applying to a CFAR workshop; and if you have been, consider mentioning it to people who fit said description.
4. If you're interested in joining us for the long haul, we're currently looking to hire a sales manager, a community manager, and an office assistant (funding permitting). We've identified these three roles as the highest-impact additions to the CFAR staff, and are eager to hear from enthusiastic and qualified candidates.

This is the mission; these are the steps. CFAR has made substantial progress on building a talent pipeline for clear thinkers and world changers, in large part thanks to generous contributions of time, money, energy, and insight from people like you. We'd like to see a world where this goal has been achieved, and your support is what gets us there. Thanks for reading; do send us any thoughts; and do please consider donating now.

Results of a One-Year Longitudinal Study of CFAR Alumni

By Dan from CFAR

Introduction

When someone comes to a CFAR workshop, and then goes back home, what is different for them one year later? What changes are there to their life, to how they think, to how they act?

CFAR would like to have an answer to this question (as would many other people). One method that we have been using to gather relevant data is a longitudinal study, comparing participants' survey responses from shortly before their workshop with their survey responses approximately one year later. This post summarizes what we have learned thus far, based on data from 135 people who attended workshops from February 2014 to April 2015 and completed both surveys.

The survey questions can be loosely categorized into four broad areas:

1. **Well-being**: On the whole, is the participant's life going better than it was before the workshop?
2. **Personality**: Have there been changes on personality dimensions which seem likely to be associated with increased rationality?
3. **Behaviors**: Have there been increases in rationality-related skills, habits, or other behavioral tendencies?
4. **Productivity**: Is the participant working more effectively at their job or other projects?

We chose to measure these four areas because they represent part of what CFAR hopes that its workshops accomplish, they are areas where many workshop participants would like to see changes, and they are relatively tractable to measure on a survey. There are other areas where CFAR would like to have an effect, including people's epistemics and their impact on the world, which were not a focus of this study.

We relied heavily on existing measures which have been validated and used by psychology researchers, especially in the areas of well-being and personality. These measures typically are not a perfect match for what we care about, but we expected them to be sufficiently correlated with what we care about for them to be worth using.

We found significant increases in variables in all 4 areas. A partial summary:

Well-being: increases in happiness and life satisfaction, especially in the work domain (but no significant change in life satisfaction in the social domain)

Personality: increases in general self-efficacy, emotional stability, conscientiousness, and extraversion (but no significant change in growth mindset or openness to experience)

Behaviors: increased rate of acquisition of useful techniques, emotions experienced as more helpful & less of a hindrance (but no significant change on measures of cognitive biases or useful conversations)

Productivity: increases in motivation while working and effective approaches to pursuing projects (but no significant change in income or number of hours worked)

The rest of this post is organized into three main sections. The first section describes our methodology in more detail, including the reasoning behind the longitudinal design and some information on the sample. The second section gives the results of the research, including the variables that showed an effect and the ones that did not; the results are summarized in a table at the end of that section. The third section discusses four major methodological concerns—the use of self-report measures (where respondents might just give the answer that sounds good), attrition (some people who took the pre-survey did not complete the post-survey), other sources of personal growth (people might have improved over time without attending the CFAR workshop), and regression to the mean (people may have changed after the workshop simply because they came to the workshop at an unusually high or low point)—and attempts to evaluate the extent to which these four issues may have influenced the results.

Methodology

This study uses a longitudinal design. Everyone who attended a four-day CFAR workshop between February 2014 and April 2015 was asked to complete a pre-workshop survey about one week prior to the workshop, and everyone who completed the pre-workshop survey was asked to take a post-workshop survey about one year later (with the timing of the post-workshop survey varying somewhat for logistical reasons). This allowed us to answer questions like: on average, are people one year after the workshop happier than they were before the workshop, or less happy, or equally happy? The null hypothesis is that they are equally happy, and statistical tests can check whether the average change in happiness is significantly different from zero.

We did conduct a small randomized controlled trial in 2012, but for our current study we chose not to include a control group. Using randomized admissions for our current study would have been expensive because it requires finding people who would like to attend a CFAR workshop and then preventing them from coming for a year. The decision to instead use a longitudinal design made it feasible to have a much larger sample size. The lack of a control group raises methodological concerns (which are discussed in more detail in the final section of this post, along with methodological concerns which would be present even with a randomized control group), but we nonetheless consider these data to be useful evidence on causal questions about the effects of the workshop.

It is worth noting that causal effects of the workshop could happen by many different pathways. For example, a participant might learn a useful technique in class, have an insight during a conversation with other participants, make a new friend, make changes shortly after the workshop which have knock-on effects, later volunteer for

CFAR, or have a shift in self-image simply from having attended an “applied rationality workshop.” The effects measured by this study (and the effects of CFAR that we care about) include all of these causal pathways, and not merely the effects that follow directly from learning CFAR content.

Some numbers on our sample: 196 people completed a pre-workshop survey shortly before attending a CFAR workshop and were asked to complete a post-workshop survey approximately one year later. 135 of them did take the post-workshop survey, a 69% response rate (though the sample size for most questions is between 122 and 132 because of skipped questions and people who started the post-workshop survey but did not finish it). On average, the post-workshop survey was taken 361 days after the pre-workshop survey (SD = 104, range of 190-564). The average age of participants on the pre-workshop survey was 26.8 years old (SD = 5.8, range of 18-43).

Results

The results here are broken down into 4 categories: well-being, personality, behaviors, and productivity. We report the results on all of the questions which were included on the survey as outcome measures.[1]

Effect sizes are given as the standardized mean difference using the standard deviation on the pre-workshop survey. An effect size of d = 0.3, for example, means that if you took a person from the post-workshop group and put them with all of the people in the pre-workshop group, you would expect them to be 0.3 standard deviations above average in that group (which is the difference between the 50th percentile and the 62nd percentile in a normal distribution). All variables are coded such that a positive effect size indicates an improvement from pre-workshop to post-workshop.

Most statistical tests used here are paired t-tests. A paired t-test takes the difference score for each participant (post-workshop score minus pre-workshop score) and then tests if the data are consistent with the null hypothesis of a difference score of zero. Results of t-tests are reported along with the degrees of freedom (which is one less than the sample size) and the test's p-value. We use a .05 threshold for statistical significance using two-tailed tests, and report effects as being statistically significant at the .05, .01, or .001 level. Effects with p < .10 are reported as nonsignificant trends, while effects with p > .10 are reported as “no change.”

Well-Being

The CFAR workshop is not explicitly targeted at making people happier, but one might expect an effective applied rationality workshop to increase its participants' well-being. Well-being can be thought of as an extremely common and general goal, or as good feelings that typically result from success at accomplishing one's goals, or as an assessment of how well one is progressing at one's goals. Thus, we considered it worth including several measures of well-being on the workshop survey.

Happiness

The Subjective Happiness Scale is a 4-item self-report measure developed by Lyubomirsky and Lepper (1999). Happiness on this scale increased by d = 0.19 (t(128) = 3.11, p < .01).

Life Satisfaction

Life satisfaction was measured using the single question “How satisfied are you with your life as a whole?” This question, and slight variants, are commonly used in wide-scale survey research. Life satisfaction increased by d = 0.17 (t(131) = 2.08, p < .05).

Domain-Specific Life Satisfaction

After the general life satisfaction question, participants were asked to rate their life satisfaction in three domains using questions which we created: “How satisfied are you with how your life is going in each of the following domains?”

In the “romantic relationships” domain, life satisfaction increased by d = 0.15 (t(130) = 2.19, p < .05).[2]

In the “work / school / career” domain, life satisfaction increased by d = 0.36 (t(131) = 3.96, p < .001).

In the “friendships / non-romantic social life” domain, life satisfaction did not change (d = 0.11, t(131) = 1.32, p = 0.19).

Social Support

While not strictly a measure of well-being, the amount of social support that a person has is closely related to their quality of life in the social domain. We created a 3-item metric of social support, which asked participants to estimate 1) the approximate number of people who they interacted with in the past week, 2) the number of people who they would be willing to confide in about something personal, and 3) the number of people who would let them crash at their place if they needed somewhere to stay. Their numerical responses were then capped at 300, log transformed, and averaged into a single measure of social support.

There was no change on this measure of social support (d = 0.11, t(122) = 1.65, p = .10).

Stuckness

Participants were asked to what extent they agreed with the statement “I feel like my life is stuck.” This question (which we created) can be interpreted as a measure of something closely related to life satisfaction, with a focus on how one's life is moving rather than on its current state.[3] Feelings of stuckness decreased by d = 0.31, t(128) = 3.02, p < .01 (with a positive effect size indicating a reduction in stuckness).

Summary of Results on Well-Being

There were significant improvements on all 3 general measures of well-being: happiness, life satisfaction, and stuckness. There were also significant improvements on domain-specific life satisfaction in 2 of the 3 domains (romantic and work). There was no change in social life satisfaction or social support. The largest effects were the increase in life satisfaction in the work domain and the decrease in feeling like one's life is stuck.

Personality

Advancing in the art of rationality involves shifts to how one thinks. There are reasons to suspect that this includes shifts along several of the dimensions that are measured by existing personality scales.

We had prior reason to suspect that the CFAR workshop had an effect on three personality variables: general self-efficacy, emotional stability, and growth mindset. That makes this survey something like a conceptual replication of those effects.[4]

General Self-Efficacy

General self-efficacy is the tendency to see problems as solvable, and to see yourself as capable of succeeding at the challenges that you face. One form that this attitude can take which is taught at CFAR workshops: if you notice an important problem, spend 5 minutes (by an actual clock) attempting to solve it.

General self-efficacy was measured using Chen and colleagues' (2001) 8-item New General Self-Efficacy Scale (sample item: “I will be able to successfully overcome many challenges”).

General self-efficacy increased by d = 0.16 (t(122) = 1.99, p < .05).

Growth Mindset

Growth mindset is the tendency to see oneself as malleable and capable of improving in important ways. In some respects, it seems to be an internal analogue of general self-efficacy: I am capable of accomplishing things even if that depends on acquiring capabilities that I do not have yet. A 4-item scale of growth mindset was taken from

Carol Dweck's book *Mindset* (sample item: “No matter what kind of person you are, you can always change substantially”).

There was no change in growth mindset (d = 0.07, t(127) = 0.79, p = .43).

This effect size (d = 0.07) is compatible with both a true effect of zero and a true effect of d = 0.21, which was the effect size that we found on the 2013 LW census, which makes this null result somewhat difficult to interpret. The average growth mindset score before the workshop (3.04 on a 1-4 scale) was already higher than the average growth mindset score on the 2013 LW census among people who had attended a CFAR workshop (2.86), which is some evidence in favor of the model where growth mindset makes people more likely to attend a CFAR workshop.

Big Five

The Big Five personality model is widely used by psychologists to study broad tendencies in people's personality. Most of the five factors have a plausible link to rationality. In declining order of the strength of that link (in my subjective opinion): emotional stability (sometimes called “neuroticism” when scored in the opposite direction) reflects resilience to stress and a tendency to not suffer from unhelpful negative emotions. It (along with general self-efficacy) is included among the 4 core self-evaluations that some psychologists consider to be fundamental traits which underlie a person's ability to act effectively in the world. Conscientiousness is closely related to the ability to reliably get things done. Openness to experience is a broad factor which includes aesthetic aspects, but is also associated with intellectual curiosity and inventiveness. Extraversion seems less core to rationality, but does seem related to having a high level of social fluency which is useful for many purposes (and which is an object-level skill that gets some training at CFAR's intensive four-day workshop). Lastly, agreeableness is related to the ability or inclination to cooperate with others, though there are also some rationality skills (such as avoiding confirmation bias and groupthink) which seem related to disagreeableness.

The 44-item Big Five Inventory (John & Srivastava, 1999) was used to measure personality on the Big Five factors.

Emotional stability increased by d = 0.13 (t(126) = 2.71, p < .01). This is consistent with our finding on the 2012 RCT.

Conscientiousness increased by d = 0.24 (t(128) = 4.38, p < .001).

Openness to experience did not change (d = 0.01, t(127) = 0.09, p = .93). Openness is sometimes separated into two facets, openness to ideas (which seems more relevant to rationality) and openness to aesthetics (which seems less relevant). There was also no change in openness to ideas (d = 0.06, t(127) = 0.87, p = .39).

Extraversion increased by d = 0.12 (t(126) = 2.82, p < .01).

Agreeableness did not change (d = 0.09, t(127) = 1.55, p = .12).

Socially Prescribed Perfectionism

Socially prescribed perfectionism is a form of perfectionism that is motivated by a desire to meet other people's extremely high expectations. Powers, Koestner, and Topciu (2005) found that people high in socially prescribed perfectionism benefited less from implementation intentions, a technique that is taught at CFAR workshops.

We included a 5-item measure of socially prescribed perfectionism, adapted from the Multidimensional Perfectionism Scale (sample item: "Anything that I do that is less than excellent will be seen as poor work by those around me"). This measure was primarily intended to investigate whether people high in socially prescribed perfectionism would respond differently to the workshop, but it would also be interesting to see if the workshop led to a reduction in socially prescribed perfectionism.

There was no change in socially prescribed perfectionism, d = -0.07, t(127) = -0.76, p = .45 (with a negative effective size indicating an increase in socially prescribed perfectionism).

Summary of Results on Personality

Replicating our 2012 RCT, we found significant increases in general self-efficacy and emotional stability. There were also significant increases in conscientiousness and extraversion. Unlike in the 2013 Less Wrong census results, there was no effect on growth mindset. There was also no change in openness to experience (or openness to ideas), agreeableness, or socially prescribed perfectionism.

Behaviors

Many of the changes that one would hope to see as a result of an applied rationality workshop are changes in habits, skills, or other behavioral tendencies, rather than changes in personality.

We included measures of a small subset of these changes: a self-report measure of the rate at which a person's toolkit of useful techniques is growing, a self-report measure of three useful conversational behaviors, performance measures of four cognitive biases or reasoning errors, and a self-report measure of whether a person relates to their emotions in a way that helps towards their goals.

Technique Acquisition Rate

The process of improvement in “applied rationality” involves acquiring new useful thinking skills, habits, and so on. As a self-report measure of their rate of progress in this process, participants were asked “On average, about how often do you find another technique or approach that *successfully helps you at* being more rational / more productive / happier / having better social relationships / having more accurate beliefs / etc.?” There were 8 response options ranging from “several times per day” to “once per year or less”.

This question was preceded by two similar questions, one which asked how often participants “read or hear about” new techniques and one which asked how often then “try out” new techniques. These other two questions helped set the context for the participant, and they allow some exploration of the process behind successful technique acquisition.

This was a conceptual replication of results of the 2013 Less Wrong census, which found that CFAR alumni reported acquiring a new technique every 81 days[5], which was significantly more often than the other LWers who acquired one every 154 days.

The rate of technique acquisition increased from once every 98 days pre-workshop to once every 59 days post-workshop, d = 0.34, t(128) = 4.53, $p < .001$.

Looking at the process of technique acquisition (assuming a simple leaky pipeline model of “hear about” → “try” → “successfully acquire”):

- There was no change in hearing about new techniques (once per 5.7 days pre-workshop vs. once per 6.7 days post-workshop, t(129) = -1.20, p = .23).
- People tried more techniques after the workshop (once per 22 days) than before (once per 30 days), t(128) = 2.24, $p < .05$.
- Success rate (acquired techniques per tried technique) increased from 31% pre-workshop to 39% post-workshop, t(127) = 2.23, $p < .05$.

These results differ somewhat from the 2013 Less Wrong census, which found that CFAR alumni tried more techniques but did not have a higher success rate.

Use of Conversations

Recognizing your flaws, identifying opportunities for improvement, and strategizing about how to act more effectively are three examples of rationality-related behaviors that often seem to benefit from talking to other people. In order to measure the extent to which people were making use of conversations for purposes like these, we asked participants these three questions:

- Can you recall a specific example from the last week when: You had a conversation with someone about whether a specific thought, feeling, or action of yours was influenced by cognitive biases or similar forms of irrationality?
- Can you recall a specific example from the last week when: A friend told you about a trait or behavior of yours that they thought you could improve upon?
- In the last month, about how many conversations have you had about specific strategies for becoming more effective at some work-related or personal task?

For the first two questions, “Yes” was coded as 1 and “No” was coded as 0. For the third question, numerical responses were capped at 50 and log transformed.

There was a nonsignificant trend towards an increase in conversations about one's own biases, from 56% to 64% (d = 0.17, t(128) = 1.78, p = .08). There was no change in conversations about traits that could be improved (d = -0.08, t(128) = -0.69, p =

.49) and no change in the number of strategic conversations (d = -0.02, t(128) = -0.26, p = .80).

When the three questions were combined into a single scale (by standardizing each item to have a standard deviation of one and then averaging), there was also no change (d = 0.02, t(126) = 0.19, p = .85).

Cognitive Biases

The survey included measures of four cognitive biases or reasoning errors: calibration, anchoring, framing effects (involving choice vs. matching question formats), and disjunctive reasoning.

Because of concerns about practice effects if participants responded to the same questions twice, each of these questions was given to half of participants on the pre-survey and to the other half on the post-survey. Unfortunately, this reduced the statistical power to a point where we were unable to detect three of these biases in our sample.

On the measures of calibration, anchoring, and framing effects we did not find a statistically significant bias among the pre-survey group or among the complete set of participants (collapsed across the pre- vs. post-workshop variable). Unsurprisingly, there was also no significant change between the pre-workshop group and the post-workshop group. Thus, these measures are not discussed in detail here.

As a measure of disjunctive reasoning, we used a question from Toplak and Stanovich (2002): “Jack is looking at Anne, but Anne is looking at George. Jack is married but George is not. Is a married person looking at an unmarried person?” Response options were “Yes,” “No,” and “Cannot be determined.”

This is a disjunctive reasoning question because it requires considering multiple scenarios. Giving the correct answer (“Yes”) requires separately considering the case where Anne is married (and looking at unmarried George) and the case where Anne is unmarried (and being looked at by married Jack).

Including both the pre-workshop and post-workshop groups, only 39% of participants answered this question correctly. Using a two-sample t-test, there was not a significant difference between the pre-workshop group (34% correct) and the post-workshop group (43% correct), d = 0.19 (t(120) = 1.01, p = .31).

For comparison, 14% of participants in Toplak and Stanovich (2002) answered correctly and 46% of Less Wrongers answered correctly on the 2012 Less Wrong census.

Emotions Help Rather Than Hinder

We created a single item measure of how participants relate to their emotions: “When you feel emotions, do they mostly help or hinder you in pursuing your goals?” The CFAR workshop emphasizes the value of emotions as sources of data and motivation, and includes techniques for increasing the alignment between one's emotions, goals, and behavior. Research on emotion regulation provides some of the theoretical background for this approach, with the findings that it generally is not helpful to suppress emotions and it often is helpful to reframe situations so that they elicit different emotions (Gross, 2002).

There was an increase in the extent to which participants evaluated their emotions as helping them rather than hindering them, d = 0.41 (t(129) = 4.13, p < .001).

Summary of Results on Behaviors

There was no sign of a change in cognitive biases. However, this study appears to have been underpowered for measuring cognitive biases in this population, as the sample as a whole did not show statistically significant effects of miscalibration, anchoring, or framing (choice vs. matching).

There was a strong increase in the rate at which participants reported acquiring new useful techniques, similar in size to the correlational result found on the 2013 Less Wrong census. Participants also became more likely to experience their emotions as helping rather than hindering their goals. However, there was no change in people's tendency to engage in three useful forms of conversation.

Productivity

Working productively is one area of life where techniques covered at the CFAR workshop are readily applicable, and most participants choose to use some productivity-related issues from their life when they practice applying the techniques at the workshop. Productivity is also relatively amenable to measurement.

The effects reported above on satisfaction with life in the domain of work/school/career (d = 0.36, p < .001), general self-efficacy (d = 0.16, p < .05), and conscientiousness (d = 0.24, p < .001) provide some evidence of improvement in the area of productivity. Additionally, the workshop survey included self-report measures (which we created) of work hours, efficiency and motivation during those work hours, income, and effective approaches to working on projects.

Work Hours

Participants were asked how many hours they spent “doing productive work” yesterday (or, if yesterday was not a workday, on their most recent workday). It is not entirely clear if an increase in work hours would be a good thing; in large part this question was asked to set the context for the next two questions.

There was no change in number of hours worked, d = -0.10, t(128) = -1.00, p = .32. [6]

Work Efficiency

Participants were asked to rate the efficiency of the time that they spent working yesterday, on a scale “between 100% efficiency, where you are working as productively as you're capable of, and 0% efficiency, where you are not getting anything done” which was provided in 10% increments.

There was a nonsignificant trend towards an increase in self-reported work efficiency, from 65% efficiency pre-workshop to 69% efficiency post-workshop (d = 0.21, t(127) = 1.91, p = .06).

Work Motivation

Participants were asked to report for how many of their work hours yesterday they felt physically motivated to do the task at hand, meaning that “the thing that you were doing was the thing that you felt like doing at that moment.”

We divided this number (which participants gave as a number of hours) by the number of hours that they worked, to give a number on a 0-1 scale.

The proportion of their work time during which they felt motivated to work on the task at hand increased from 56% to 63% (d = 0.24, t(121) = 2.30, p < .05).

This question and the work efficiency question can both be considered measures of the quality of one's work hours. If both questions are scaled from 0-1 and they are averaged together to create a single measure, the increase on that scale is d = 0.28 (t(121) = 2.56, p < .05).

Income

Participants were asked to report their income over the past year (the 365 days directly prior to completing the survey). Participants were given the option of selecting their currency, and incomes reported in non-US currencies were converted to US dollars using the exchange rates on November 30, 2015. Reported incomes were then log transformed using the equation log(income + $10,000) to reduce skew while preventing low incomes (which could have a variety of causes, such as being a student) from being given undue weight.

There was no change in income (d = 0.05, t(118) = 0.94, p = .35).[7]

Effective Approaches to Working on Projects

We created a 4-item self-report measure of people's tendency to use effective approaches when working on projects. Specifically, participants were asked to rate how regularly they acted in accordance with the following statements (on a 1-6 scale from “Almost never” to “Almost always”):

- When I decide that I want to do something (like doing a project, developing a new regular practice, or changing some part of my lifestyle), I ...
    1. ...plan out what specific tasks I will need to do to accomplish it.
    2. ...try to think in advance about what obstacles I might face, and how I can get past them.
    3. ...seek out information about other people who have attempted similar projects to learn about what they did.
    4. ...end up getting it done.

The four items were averaged into a single measure of effective approaches to projects.

The use of effective approaches to projects increased by d = 0.45 (t(128) = 6.19, p < .001).[8]

Summary of Results on Productivity

There was no change in income over this one year time period, or in the number of hours worked. There was a nonsignificant trend towards participants reporting that yesterday they worked closer to the highest level of productivity that they are capable of. There were significant increases on the other measures of what participants did while they were working (or otherwise involved in important projects). Participants reported spending a larger fraction of their working time with the feeling that the thing that they were doing was the thing that they felt like doing at that moment, using more effective approaches to working on projects, and getting more of their projects done.

These results are consistent with the findings reported earlier on increased conscientiousness, increased general self-efficacy, and higher satisfaction with one's life in the domain of work/school/career.

General Summary

The table below summarizes the study results:

Category	Measure	Effect Size (d)
Well-being	Subjective Happiness Scale	0.19**
Well-being	Life Satisfaction	0.17*
Well-being	Life Satisfaction: Romantic	0.15*
Well-being	Life Satisfaction: Work/School/Career	0.36***
Well-being	Life Satisfaction: Social	0.11
Well-being	Social Support	0.11
Well-being	Stuckness (R)	0.31**

Personality	General Self-Efficacy	0.16*
Personality	Growth Mindset	0.07
Personality	Emotional Stability	0.13**
Personality	Conscientiousness	0.24***
Personality	Openness to Experience	0.01
Personality	Extraversion	0.12**
Personality	Agreeableness	0.09
Personality	Socially Prescribed Perfectionism (R)	-0.07
Behaviors	Technique Acquisition Rate	0.34***
Behaviors	Use of Conversations	0.02
Behaviors	Cognitive Biases: Disjunctive Reasoning	0.19
Behaviors	Emotions Help Rather Than Hinder	0.41***
Productivity	Work Hours	-0.10
Productivity	Work Efficiency	0.21†
Productivity	Work Motivation	0.24*
Productivity	Income	0.05
Productivity	Effective Approaches to Working on Projects	0.45***

Table 1: Summary of Effect Sizes. All variables are coded such that positive numbers indicate an improvement from pre-workshop to post-workshop, with (R) indicating that this involved reverse scoring. Effect size is the standardized mean difference using the pre-workshop standard deviation. † p<.10, * p<.05, ** p<.01, *** p<.001.

Methodological Concerns

We found statistically significant effects on many of the variables that we measured, but it is important to consider whether these results are likely to be due to a causal effect of the workshop or if there are methodological flaws which would cause these effects to be found even in the absence of any causal effect of the workshop.

In this section, we consider four methodological concerns that seem like especially plausible sources of bias in our results:

- **Self-Report and Socially Desirable Responding**: High scores on self-report measures may reflect participants giving the response that they see as desirable rather than them actually doing well on the construct that we are attempting to measure.
- **Attrition**: If the people who benefited the least from the workshop did not respond to the post-workshop survey, then the effects measured among the people who did take the post-survey could be inflated.
- **Other Sources of Personal Growth**: If the participants are improving over time for other reasons, then improvement from the pre-survey to the post-survey could be unrelated to the workshop.

- **Selection Effects and Regression to the Mean**: If participants take the pre-survey at an unusual time (such as when their life feels especially stuck), then improvements on the post-survey could reflect regression to the mean.

This section discusses the extent to which these four concerns are likely to have influenced our results, and includes several supplementary analyses that cast some light on that question.

Self-Report and Socially Desirable Responding

One clear concern about this research is the degree to which it relies on self-report measures. Further, most of these measures involved questions about relatively broad tendencies (rather than asking about concrete, observable behaviors) and asked for responses on a scale where it was easy to identify one end of the scale as being better than the other.

Many of the measures that we used have been validated by the existing research literature. For example, the Subjective Happiness Scale is correlated with reports from people's friends of how happy they are (Lyubomirsky & Lepper, 1999). Thus, even if there is some tendency for people to respond based on how desirable the options are, we would also expect their answers to reflect the underlying construct which the scale is attempting to measure.

The relevant question, in assessing these results, is to what extent the changes between the pre-survey and the post-survey reflect changes in the underlying construct and to what extent they reflect changes in people's tendency to give the desirable response. Because the same person is answering the same questions on the pre- and post-surveys, one might expect the tendency to give desirable responses to be similar on both surveys, and therefore to cancel out.

There is some concern that people would be especially motivated to give high responses on the post-survey, if they had the opinion that they benefited from the workshop and wanted to validate this opinion (though this is made somewhat less likely by the fact that they were asked to report on their current tendencies, and not to make comparisons with their pre-workshop selves). However, people also might be especially motivated to give high responses on the pre-survey, as they were about to attend a rationality workshop and were answering questions about themselves to the people who were going to run that workshop.

The measures that seem most likely to show an increase after the workshop due to socially desirable responding are those where participants' opinions about what response is desirable changed because of the workshop. For example, my impression is that CFAR workshops increase participants' tendency to believe that it is a good idea to “try to think in advance about what obstacles I might face, and how I can get past them” and that it is desirable to arrange it so that your emotions help (rather than hinder) you in pursuing your goals. Thus, it seems reasonable to put less stock in this study as evidence that people actually changed the extent that they plan for obstacles in advance, and are helped rather than hindered by their emotions, compared to other measures which showed similar changes.

The experimental design which is typically used to reduce this concern is to use a control group which receives a placebo treatment. This design is infeasible given that our intervention is an intensive workshop. A randomized control group which received no intervention would not provide a similar advantage, since we are most concerned about socially desirable responding from people who have attended the workshop.

Another approach to reducing this concern is to rely on peer report rather than self report. We have been conducting a peer survey of CFAR workshop participants, which involves sending surveys about the participant to 2 of their friends, both before the workshop and again approximately one year later. We are in the final stages of data collection on those surveys, and expect to begin the data analysis later this month.

Attrition

People who begin a study but do not complete it are a concern for any research with a longitudinal design, because the people who complete the study (and thereby provide researchers with data that they can use) are not necessarily representative of the people who started the study. That is, attrition creates the possibility of nonresponse bias. For example, some drug treatment programs report an extremely high success rate among people who finish the program only because the people who relapse almost all drop out of the program.

Unlike in the case of drug relapsers, a person who did not benefit from the CFAR workshop may still be fairly likely to respond to the post-workshop CFAR survey. A 69% response rate is within the norm for this type of survey-based longitudinal research, and nonresponse can occur for a variety of reasons (e.g., being busy, disliking the hassle of filling out a half-hour survey). Still, the 31% attrition rate does raise some questions about what results would have been found if the whole pre-survey sample of 196 people had completed the post-survey. Taking the results at face value implicitly assumes that there is no difference between those who did complete the post-survey and those who did not, which seems implausible, since we would generally expect some correlation between how much a person benefited from CFAR and how willing they are to complete a survey for CFAR.

We could make the pessimistic assumption that the nonrespondents had zero net change, which would imply that the actual effect sizes are only 69% as large as the effect sizes reported here (since there was a 69% response rate). I suspect that this is too pessimistic, though it is worth noting that it still is not a lower bound—presumably there are some people who got worse over the course of a year, and one might expect them to be especially unlikely to take the post-survey.

One way to estimate the size of the effect of attrition is to look at the order in which the people in each workshop cohort took the post-survey. It seems likely that nonrespondents have more in common with the last people to take the survey (who received multiple reminders) than with the first people to take the survey immediately after it was sent out. Below is a graph of effect size vs. response order.

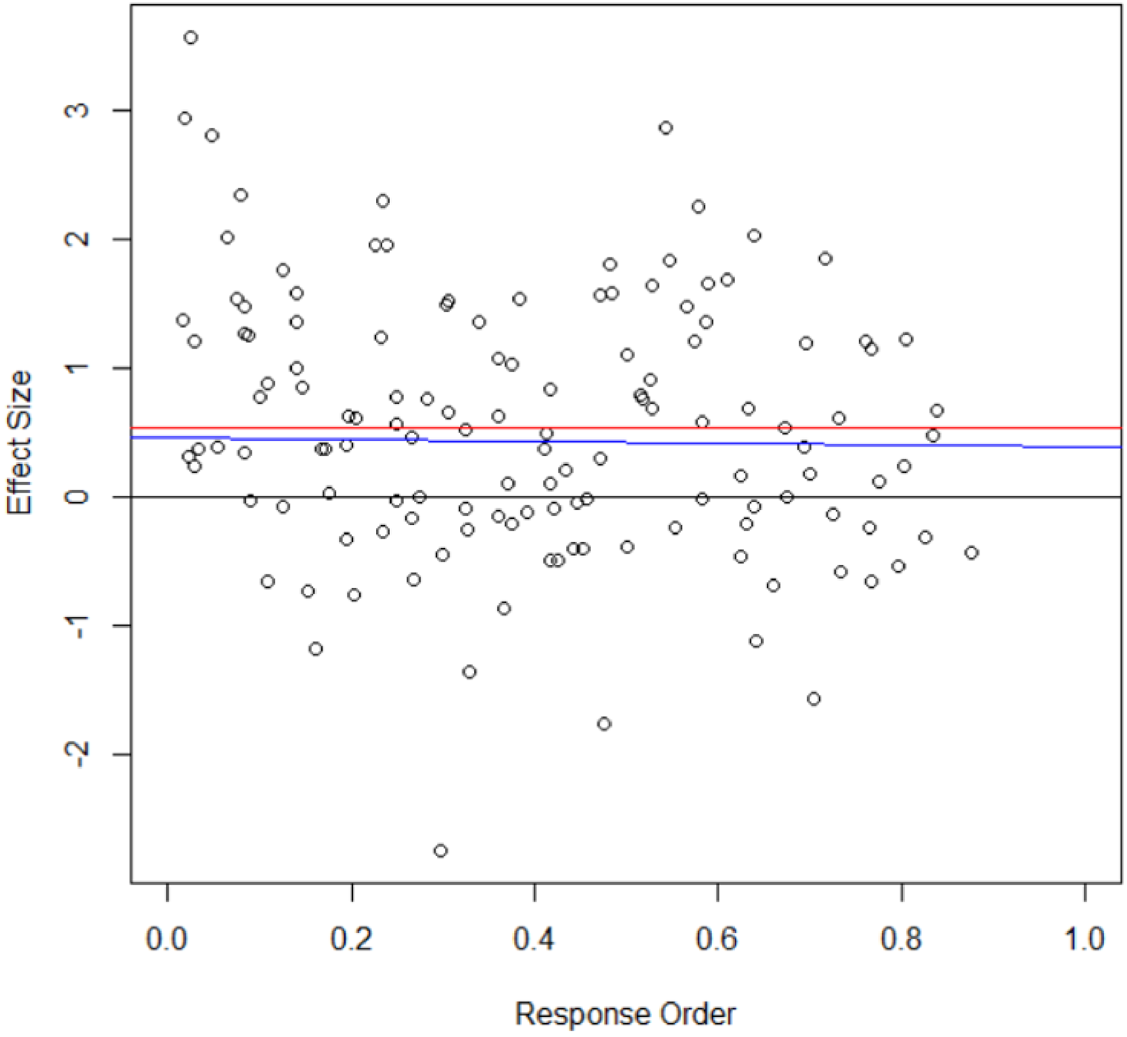


Figure 1: Effect Size by Response Order. The graph includes the sample mean (in red) and the best fit regression line for the participants with Response Order of at least 0.10 (in blue).

Effect size is based on a composite measure of the variables that had statistically significant effects (where 0 = no change from pre-workshop and SD = 1).[9] Response order is scaled from 0 to 1, where 0 is the first person from each workshop to complete the survey and 1 would be the last person from the workshop if everyone completed the survey (the blank space on the right side of the graph is a result of having a response rate below 100%). The red line shows the mean among everyone who took the survey (0.54 standard deviations above zero), and the blue line is a best-fit linear regression line when excluding the first 10% of respondents.

The apparent pattern is that the effect size was higher among the first 10% of respondents, and then relatively flat (or trending slightly downward) among the rest of respondents. If we extrapolate the slight downward trend to fill in the missing data from nonrespondents (by fitting a linear regression line to everyone except the first 10%, as shown in blue), the implied effect size is 0.51 for the whole group of 196 participants who completed the pre-workshop survey, which is only slightly less than the observed 0.54 effect size for the 135 participants who took the post-workshop survey.[10]

An alternative way to estimate the effect of attrition is to pick a subset of nonrespondents and make an extra effort to cause them to complete the post-survey. We did this as well, focusing on the two most recent workshops (January 2015 and April 2015). These workshops already had a 77% response rate (probably because of improvements to our method of soliciting responses), and after extra cajoling of the nonrespondents we were able to increase the response rate to 88% (42/48). The average effect size for participants of the January 2015 and April 2015 workshops (with 88% response rate) was 0.54, which is the same as the 0.54 effect size for all 135 respondents (with 69% response rate).

It appears that attrition probably inflated our effect sizes by only a modest amount.

Other Sources of Personal Growth

The primary weakness of longitudinal studies, compared with studies that include a control group, is that they cannot easily distinguish changes due to the intervention from changes that would have occurred over time even without the intervention. It is possible that the people who attended the CFAR workshops would have experienced increases in self-efficacy, well-being, emotional stability, and other variables regardless of whether or not they attended a workshop.

Most of the variables that showed increases among CFAR workshop participants do not tend to increase over time among this age group. Happiness (Stone et al., 2010), life satisfaction (Qu & de Vaus, 2015; Stone et al., 2010), emotional stability (Soto et al., 2011; Srivastava et al., 2003), and extraversion (Soto et al., 2011; Srivastava et al., 2003) do not increase on average over a person's 20s. Conscientiousness does increase (Soto et al., 2011; Srivastava et al., 2003), but only by about d = 0.03 per year, which is much less than the d = 0.24 increase that we found about one year after the workshop.

One might suspect that people who chose to attend an applied rationality workshop are different from the general population, in a way that would involve a larger increase over time on variables like these. One way to estimate the size of this effect is to look at the correlation between age and the outcome variables on the pre-survey. If CFAR participants are coming from a population whose self-efficacy is increasing with age (for example), the older people in that group should have higher self-efficacy than the younger people. (Though this relationship is attenuated by the fact that workshop attendance involves a filter, and that age is an imperfect proxy for the amount of time that a person has spent heavily focused on personal growth).

Age was essentially uncorrelated with participants' overall score on the variables that increased after the workshop, r(132) = .07, p = .41.[11] Three of the twelve pre-workshop variables did have a statistically significant correlation with age: quality of hours worked (t(120) = 2.27, p < .05), general self-efficacy (t(131) = 2.26, p < .05), and conscientiousness (t(134) = 2.03, p < .05). Each additional year of age predicted a d = 0.03 increase in quality of hours worked, a d = 0.03 increase in general self-efficacy, and a d = 0.03 increase in conscientiousness.

Another way to get evidence on how much our population of participants is improving over time (independent of the workshop) involves looking at how much time passed between the pre-survey and the post-survey. Conveniently, there was a large amount of variance in this time gap (mean = 361 days, SD = 104, range of 190-564) because we did not make it a priority to send out the “one year” followup survey exactly one year after the workshop, and we also recently chose to reduce the length of the time gap to 6 months. If people are improving over time for reasons unrelated to the workshop, then we would expect a much larger improvement from people who took the post-survey 15 months after the pre-survey than from people who took the post-survey after only a 6 month gap. This time gap was uncorrelated with overall effect size (r(134) = -.05, p = .55) and slightly in the direction of a smaller effect size for people with a larger time gap.[12]

Including a control group in the research design would have allowed for more confident conclusions, but it seems relatively unlikely that the survey results are primarily capturing personal growth over time which would have occurred even without the workshop.

Selection Effects and Regression to the Mean

A related limitation of longitudinal research is that the study may begin at an unusual time in the participant's life, which is then followed by changes over time simply due to regression to the mean. It is difficult to study treatments for depression using a longitudinal design (without a control group), for example, because depressive episodes typically last for a period of months, so a group of participants who start out depressed can be expected to get better regardless of the treatment.

In the case of a CFAR workshop, one might worry that people tend to choose to attend a rationality workshop at an unusual time in their life (especially on the dimension of feeling good about the direction that their life is headed). One also might worry that feelings of anticipation in the week before a workshop might influence a participant's state of mind as they take the pre-survey (e.g., they might be excited or nervous). The practical aspects of preparing to go to a workshop (e.g., getting ready to travel and miss a couple days of work) might also make the week before a workshop an unusual time (especially for measures of the quantity and quality of work hours).

The first of these three sources of unusualness strikes me as the most concerning to the results taken as a whole. There is some reason to suspect that people will tend to decide to come to a CFAR workshop when they are at a relatively high point (e.g., feeling especially optimistic about what they can accomplish and motivated to acquire exciting new tools). There is also some reason to suspect that people will tend to come to a CFAR workshop at a relatively low point (e.g., feeling like their life is stuck and grasping for something that might help get them unstuck). If either of these tendencies dominates, then we might expect to find effects on many of the measures in this study (such as those related to well-being, productivity, and self-efficacy) due to regression to the mean in the absence of a causal effect of the workshop. People who come in at a relatively high point would be expected to show declines, and people who come in at a relatively low point would be expected to show improvements.

This concern is somewhat attenuated, because many CFAR workshop participants sign up months in advance of their workshop. This gives them time to regress towards their baseline state after they decide to come to a workshop and before they fill out the pre-workshop survey.

In order to collect more information on the likely direction and strength of this effect, I polled the CFAR staff on the question of whether participants tend to arrive at the CFAR at a relatively high point or a relatively low point, without telling them the reason for the poll. The average answer was slightly on the side of thinking that participants arrive at a relatively high point.[13] This result suggests that our best guess is that the effect sizes given here are relatively unaffected by regression to the mean on this dimension, or perhaps are slight underestimates, though a significant amount of uncertainty remains.

Conclusion

CFAR has described its long-term aim as creating a community of people with competence, epistemic rationality, and do-gooding who will be able to solve important problems that the world faces. This longitudinal study of workshop participants represents one attempt to try to collect information about our progress on some substeps of this long-term project.

On several different measures, we found that CFAR alumni had changed from how they were before the workshop in ways that seem related to increased competence, productivity, or personal effectiveness. These included changes on personality scales (general self-efficacy, conscientiousness), changes on reports about their behavior (quality of hours worked yesterday, technique acquisition rate, use of effective approaches to working on projects), and changes related to how they engaged with their emotions (emotional stability, experiencing emotions as helping rather than hindering them).

The overall change was substantial enough to be apparent on broad measures of happiness and life satisfaction. Life satisfaction was especially likely to increase in the domain of work/school/career, which suggests that these broad increases in well-being may be related to an increase in competence.

We are currently working on our plans for what to measure in the upcoming year, and hope to continue to gather data that will flesh out more of the picture of what impacts CFAR has on the people who become involved in its trainings.

Footnotes

[1] The survey also included several demographic questions which are not discussed here, including questions about the participant's involvement in the rationality community. Additionally, several questions were added or removed from the survey while this study was in progress, and are not reported on here. Six questions which were on the original (February 2014) version of the survey were removed beginning

with the June 2014 cohort in order to make the survey briefer, one more question was removed beginning with the November 2014 cohort, and one new question was added beginning with the January 2015 cohort.

[2] Exploratory analysis suggests that the increase in life satisfaction in the romantic domain was primarily due to an increase in the percent of people who were in a romantic relationship. The percent of people in a romantic relationship increased from 58% pre-workshop to 67% post-workshop, t(130) = 2.61, $p < .01$. Among those who were not in a relationship, life satisfaction in the romantic domain on a 1-7 scale was 2.52 pre-workshop and 2.62 post-workshop. Among those in a relationship, it was 5.63 pre-workshop and 5.64 post-workshop.

[3] “I feel like my life is stuck” is also somewhat related to self-efficacy, and in our 2012 RCT we treated it as a general self-efficacy item.

[4] The 2012 RCT found significant ($p < .01$) increases on brief measures of general self-efficacy and emotional stability. It is important to seek replication for those effects because the previous survey had a small sample size and a large number of outcomes variables, which increases the chance of a false positive and will tend to inflate the effect size on variables that are found to be significant. On the 2013 Less Wrong census, there was a significant association between growth mindset and whether a person had attended a CFAR workshop, d = 0.21 (t(1347) = 1.98, $p < .05$), but this result was only correlational and could result from people high in growth mindset being more likely to choose to attend a CFAR workshop.

[5] Statistical tests were conducted using the 1-8 scale. For ease of interpretation, results are reported here as days per technique. The means on the 1-8 scale correspond approximately to geometric means in days per technique. The results reported here for the 2013 Less Wrong census differ from the results that were posted on Less Wrong in 2014 because the results reported in 2014 controlled for other variables and the results reported here do not.

[6] On the work hours, work efficiency, and work motivation questions, we removed from the sample one participant who reported more than 24 hours of work, since their interpretation of the question differed from ours. On the work efficiency and work motivation questions, we also removed one participant who reported less than 1 hour of work, since those measures are much less meaningful with a small denominator. On the work motivation question, we also removed 2 participants who reported more motivated hours than total hours of work.

[7] We repeated the analysis of income while excluding all participants who reported that they were a student on either the pre-survey or the post-survey. The results were extremely similar (d = 0.06, t(68) = 0.75, p = .45).

[8] If analyzed separately, three of the four items on the measure of effective approaches to working on projects were statistically significant (planning specific tasks, planning for obstacles, and getting it done) while one was not (seeking information about other people). Planning for obstacles is the item which was most directly covered at the workshop, and it also had the largest effect size of the four items (d = 0.48).

[9] The overall measure of effect size included 12 variables, with a preference for including composite variables rather than individual items. These variables were: happiness, life satisfaction, domain-specific life satisfaction (an average of life satisfaction in the 3 domains), stuckness (reverse-scored), general self-efficacy,

emotional stability, conscientiousness, extraversion, technique acquisition rate, emotions help rather than hinder, quality of work hours (an average of efficiency and motivation), and effective approaches to projects. For each of these 12 measures, we created a variable representing the post-workshop minus pre-workshop difference score, and then rescaled it to have a standard deviation of 1. Missing values were coded as 0 (meaning no change from pre-survey to post-survey). These 12 variables were averaged for each participant, and the resulting variable was rescaled to have a standard deviation of 1. This is the overall effect size measure that was used. (The results are extremely similar if missing values are coded as the scale mean rather than as 0.)

[10] For the participants with response order of at least 0.10, the best fit line (shown in blue in Figure 1) is:

Effect Size = 0.46704 - 0.06985 * (Response Order)

This line implies an average effect size of 0.43 for the 90% of the original sample with response order of at least 0.10. For those with response order below 0.10, the average effect size was 1.29. The implied effect size for the whole group is therefore 0.9*0.43 + 0.1*1.29, which equals 0.51.

[11] This overall score metric was created in much the same way as the overall effect size metric described in [9], using the same 12 variables. It used the pre-workshop scores rather than the difference scores, and missing data was replaced with the sample mean rather than 0.

[12] This pattern of results (overall effect size not significantly correlated with time gap, and very weakly in the direction of smaller benefit with larger time gap) continues to hold after controlling for potential confounds: response order (described earlier), workshop order (coded as 1 for February 2014 and 9 for April 2015), and workshop attrition rate (the percent of participants from a person's workshop cohort who did not respond to the post-survey).

[13] I emailed the CFAR staff asking them to give a rating on a -10 to 10 scale, where -10 means that there is a strong tendency for participants to show up at a relatively low point, 10 means there is a strong tendency for them to show up at a relatively high point, and 0 means they tend to show up at their average level. Five staff members responded, with ratings of +3, -4, +3, -2, and +2 (average of 0.4). All five staff members also emphasized the uncertainty of their guess.

References

Chen, G., Gully, S. M., & Eden, D. (2001). Validation of a new general self-efficacy scale. *Organizational Research Methods January*, *4*, 62-83.

Dweck, C. S. (2006). *Mindset: The new psychology of success*. New York: Random House.

Gross, J. J. (2002). Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology*, *39*, 281-91.

John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (Vol. 2, pp. 102–138). New York: Guilford Press.

Lyubomirsky, S., & Lepper, H. S. (1999). A measure of subjective happiness: Preliminary reliability and construct validation. *Social Indicators Research*, *46*, 137-155.

Powers, T. A., Koestner, R. & Topciu, R. A. (2005). Implementation intentions, perfectionism, and goal progress: Perhaps the road to hell is paved with good intentions. *Personality and Social Psychology Bulletin*, *31*, 902-912.

Qu, L. and de Vaus, D. (2015). Life satisfaction across life course transitions. *Australian Family Trends*, *8*.

Soto, C. J., John, O. P., Gosling, S. D. & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big Five domains and facets in a large cross-sectional sample. *Journal of Personality and Social Psychology*, *100*, 330-348.

Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology*, *84*, 1041-1053. (Includes a table of Big Five personality by age.)

Stone, A. A., Schwartz, J. E., Broderick, J. E., & Deaton, A. (2010). A snapshot of the age distribution of psychological well-being in the United States. *Proceedings of the National Academy of Sciences*, *107*, 9985-9990.

The art of grieving well

[This is one post I've written in an upcoming sequence on what I call "yin". Yin, in short, is the sub-art of giving perception of truth absolutely no resistance as it updates your implicit world-model. Said differently, it's the sub-art of subconsciously seeking out and eliminating ugh fields and also eliminating the inclination to form them in the first place. This is the first piece I wrote, and I think it stands on its own, but it probably won't be the first post in the final sequence. My plan is to flesh out the sequence and then post a guide to yin giving the proper order. I'm posting the originals on my blog , and you can view the original of this post here , but my aim is to post a final sequence here on Less Wrong.]

In this post, I'm going to talk about grief. And sorrow. And the pain of loss.

I imagine this won't be easy for you, my dear reader. And I wish I could say that I'm sorry for that.

…but I'm not.

I think there's a skill to seeing horror clearly. And I think we *need* to learn how to see horror clearly if we want to end it.

This means that in order to point at the skill, I need to also point at real horror, to show how it works.

So, I'm not sorry that I will make you uncomfortable if I succeed at conveying my thoughts here. I imagine I have to.

Instead, I'm sorry that we live in a universe where this is necessary.

If you Google around, you'll find all kinds of lists of what to say and avoid saying to a grieving person. For reasons I'll aim to make clear later on, I want to focus for a moment on some of the things *not* to say. Here are a few from Grief.com:

- "He is in a better place."
- "There is a reason for everything."
- "I know how you feel."
- "Be strong."

I can easily imagine someone saying things like this with the best of intentions. They see someone they care about who is suffering greatly, and they want to help.

But to the person who has experienced a loss, these are very unpleasant to hear. The discomfort is often pre-verbal and can be difficult to articulate, especially when in so much pain. But a fairly common theme is something like:

"*Don't heave your needs on me. I'm too tired and in too much pain to help you.*"

If you've never experienced agonizing loss, this might seem really confusing at first — which is why it seems tempting to say those things in the first place, I think. But try assuming that the grieving person sees the situation more clearly, and see if you can make sense of this reaction before reading on.

…

…

…

If you look at the bulleted statements above, there's a way of reading them that says "You're suffering. Maybe try this, to stop your suffering." There's an imposition there, telling the grieving person to add more burden to how they are in the moment. In many cases, the implicit request to stop suffering comes from the speaker's discomfort with the griever's pain, so an uncharitable (but sometimes accurate) read of those statements is "I don't like it when you hurt, so stop hurting."

Notice that the person who lost someone doesn't have to think through all this. They just see it, directly, and emotionally respond. They might not even be able to say *why* others' comments feel like impositions, but there's very little doubt that they do. It's just that social expectations take *so much energy*, and the grief is already so much to carry, that it's hard *not* to notice.

There's only energy for what really, actually matters.

And, it turns out, not much matters when you hurt that much.

I'd like to suggest that grieving is how we experience the process of a very, very deep part of our psyches becoming familiar with a painful truth. It doesn't happen only when someone dies. For instance, people go through a very similar process when mourning the loss of a romantic relationship, or when struck with an injury or illness that takes away something they hold dear (e.g., quadriplegia). I think we even see smaller versions of it when people break a precious and sentimental object, or when they fail to get a job or into a school they had really hoped for, or even sometimes when getting rid of a piece of clothing they've had for a few years.

In general, I think familiarization looks like tracing over all the facets of the thing in question until we intuitively expect what we find. I'm particularly fond of the example of arriving in a city for the first time: At first all I know is the part of the street right in front of where I'm staying. Then, as I wander around, I start to notice a few places I want to remember: the train station, a nice coffee shop, etc. After a while of exploring different alleyways, I might make a few connections and notice that the coffee shop is actually just around the corner from that nice restaurant I went to on my second night there. Eventually the city (or at least those parts of it) start to feel smaller to me, like the distances between familiar locations are shorter than I had first thought, and the areas I can easily think of now include several blocks rather than just parts of streets.

I'm under the impression that grief is doing a similar kind of rehearsal, but specifically of pain. When we lose someone or something precious to us, it hurts, and we have to practice anticipating the lack of the preciousness where it had been before. We have to familiarize ourselves with the absence.

When I watch myself grieve, I typically don't find myself just thinking "This person is gone." Instead, my grief wants me to call up specific images of recurring events — holding the person while watching a show, texting them a funny picture & getting a smiley back, etc. — and then add to that image a feeling of pain that might say "...and that will never happen again." My mind goes to the feeling of *wanting* to watch a show with that person and remembering they're not there, or knowing that if I send a text they'll never see it and won't ever respond. My mind seems to want to rehearse the pain that will happen, until it becomes familiar and known and eventually a little smaller.

I think grieving is how we experience the process of changing our emotional sense of what's true to something worse than where we started.

Unfortunately, that can feel on the inside a little like *moving to* the worse world, rather than recognizing that we're already here.

It looks to me like it's possible to *resist* grief, at least to some extent. I think people do it all the time. And I think it's an error to do so.

If I'm carrying something really heavy and it slips and drops on my foot, I'm likely to yelp. My initial instinct once I yank my foot free might be to clutch my foot and grit my teeth and swear. But in doing so, even though it *seems* I'm focusing on the pain, I think it's more accurate to say that I'm *distracting myself from* the pain. I'm too busy yelling and hopping around to really experience exactly what the pain feels like.

I could instead turn my mind to the pain, and look at it in exquisite detail. Where exactly do I feel it? Is it hot or cold? Is it throbbing or sharp or something else? What exactly is the most aversive aspect of it? This doesn't stop the experience of pain, but it does stop most of my inclination to jump and yell and get mad at myself for dropping the object in the first place.

I think the first three so-called "stages of grief" — denial, anger, and bargaining — are avoidance behaviors. They're attempts to distract oneself from the painful emotional update. Denial is like trying to focus on anything other than the hurt foot, anger is like clutching and yelling and getting mad at the situation, and bargaining is like trying to rush around and bandage the foot and clean up the blood. In each case, there's an attempt to keep the mind preoccupied so that it can't start the process of tracing the pain and letting the agonizing-but-true world come to *feel true*. It's as though there's a part of the psyche that believes it can prevent the horror from being real by avoiding coming to feel as though it's real.

The above might seem kind of abstract, so let me list a very few examples that I think do in fact apply to resisting grief:

- After a breakup, someone might refuse to talk about their ex and insist that no one around them bring up their ex. They might even start dating a lot more right away (the "rebound" phenomenon, or dismissive-avoidant dating patterns). They might insist on acting like their ex doesn't exist, for months, and show flashes of intense anger when they find a lost sweater under their bed that had belonged to the ex.
- While trying to finish a project for a major client (or an important class assignment, if a student), a person might realize that they simply don't have the time they need, and start to panic. They might pour all their time into it, even while knowing on some level that they can't finish on time, but trying desperately anyway as though to avoid looking at the inevitability of their meaningful failure.

- The homophobia of the stereotypical gay man in denial looks to me like a kind of distraction. The painful truth for him here is that he is something he thinks it is wrong to be, so either his morals or his sense of who he is must die a little. Both are agonizing, too much for him to handle, so instead he clutches his metaphorical foot and screams.

In every case, the part of the psyche driving the behavior seems to think that it can hold the horror at bay by preventing the emotional update that the horror is real. The problem is, success requires severely distorting your ability to see what is real, and also your *desire* to see what's real. This is a cognitive black hole — what I sometimes call a "metacognitive blindspot" — from which it is enormously difficult to return.

This means that if we want to see reality clearly, we have to develop some kind of skill that lets us *grieve well* — without resistance, without flinching, without screaming to the sky with declarations of war as a distraction from our pain.

We have to be willing to look directly and unwaveringly at horror.

In 2014, my marriage died.

A friend warned me that I might go through two stages of grief: one for the loss of the relationship, and one for the loss of our hoped-for future together.

She was exactly right.

The second one hit me really abruptly. I had been feeling solemn and glum since the previous night, and while riding public transit I found myself crying. Specific imagined futures — of children, of holidays, of traveling together — would come up, as though raising the parts that hurt the most and saying "See this, and wish it farewell."

The pain was *so much*. I spent most of that entire week just moving around slowly, staring off into space, mostly not caring about things like email or regular meetings.

Two things really stand out for me from that experience.

First, there were still impulses to flinch away. I wanted to cry about how the pain was *too much to bear* and curl up in a corner — but I could tell that impulse came from a different place in my psyche than the grief did. It felt *easier* to do that, like I was trading some of my pain for suffering instead and could avoid being present to my own misery. I had worked enough with grief at that point to intuit that I needed to process or digest the pain, and that this slow process would go even more slowly if I tried not to experience it. It required a choice, every moment, to keep my focus on *what hurt* rather than on *how much* it hurt or how unfair things were or any other story that decreased the pain I felt in that moment. And it was tiring to make that decision continuously.

Second, there were some things I *did* feel were important, even in that state. At the start of this post I referenced how mourners can sometimes see others' motives more plainly than those others can. What I imagine is the same thing gave me a clear sense of how much nonsense I waste my time on — how most emails don't matter, most meetings are pointless, most curriculum design thoughts amount to rearranging deck chairs on the Titanic. I also vividly saw how much nonsense I *project* about who I am and what my personal story is — including the illusions I would cast on *myself*. Things like how I thought I needed people to admire me to feel motivated, or how I felt most powerful when championing the idea of ending aging. These stories looked embarrassingly false, and I just didn't have the energy to keep lying to myself about them.

What was left, after tearing away the dross, was simple and plain and beautiful in its nakedness. I felt like I was *just me*, and there were a very few things that still really mattered. And, even while drained and mourning for the lovely future that would never be, I found myself working on those core things. I could send emails, but they had to matter, and they couldn't be full of blather. They were richly honest and plain and simply directed at making the *actually* important things happen.

It seems to me that grieving well isn't just a matter of learning to look at horror without flinching. It also lets us see through certain kinds of illusion, where we think things matter but at some level have always known they don't.

I think skillful grief can bring us more into touch with our faculty of *seeing the world plainly as we already know it to be*.

I think we, as a species, dearly need to learn to see the world clearly.

A humanity that makes global warming a politicized debate, with name-calling and suspicion of data fabrication, is a humanity that does not understand what is at stake.

A world that waits until its baby boomers are doomed to die of aging before taking aging seriously has not understood the scope of the problem and is probably still approaching it with distorted thinking.

A species that has great reason to fear human-level artificial intelligence and does not pause to seriously figure out what if anything is correct to do about it (because "that's silly" or "the Terminator is just fiction") has not understood just how easily it can go horribly wrong.

Each one of these cases is bad enough — but these are *just examples* of the result of collectively distorted thinking. We will make mistakes this bad, and possibly worse, again and again as long as we are willing to let ourselves turn our awareness away from our own pain. As long as the world feels safer to us than it actually is, we will risk obliterating everything we care about.

There is hope for immense joy in our future. We have conquered darkness before, and I think we can do so again.

But doing so requires that we see the world clearly.

And the world has devastatingly more horror in it than most people seem willing to acknowledge.

The path of clear seeing is agonizing — but that is because of the truth, not because of the path. We are in a kind of hell, and avoiding seeing that won't make it less true.

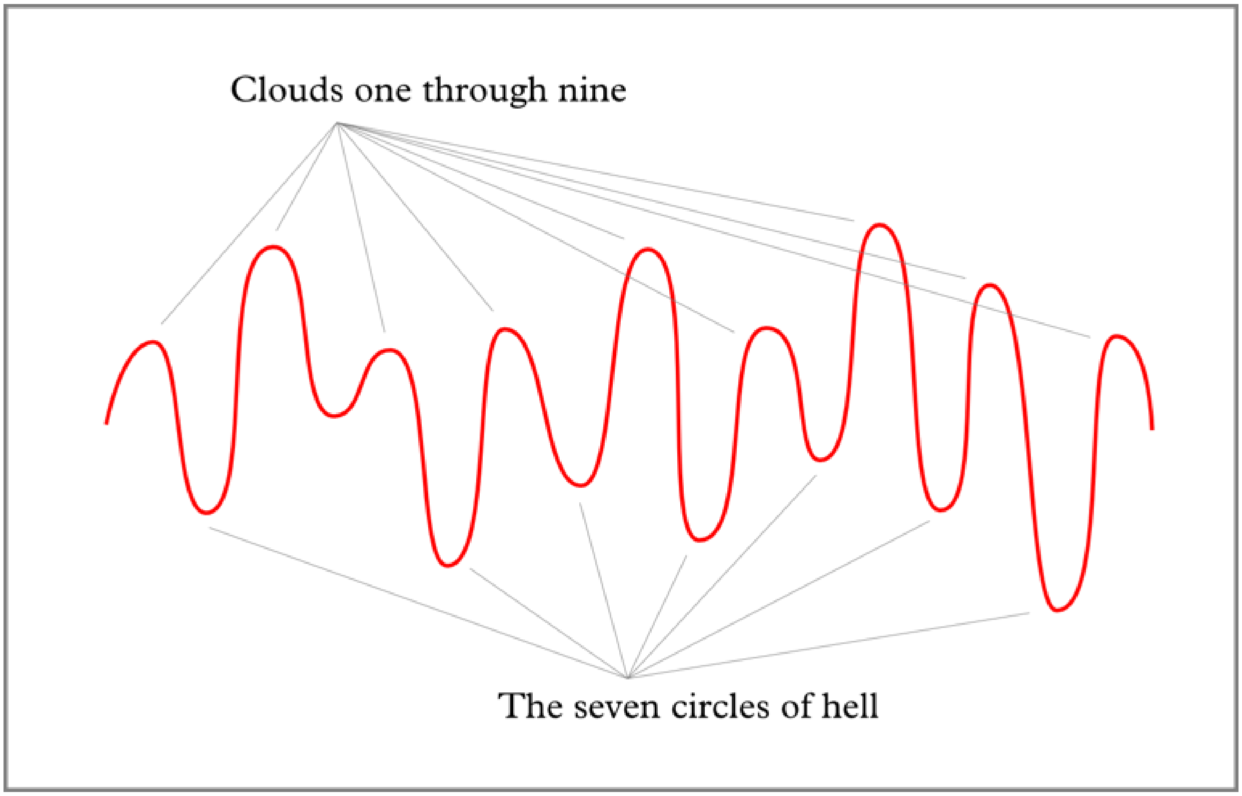
But maybe, if we see it clearly, we can do something about it.

Grieve well, and awaken.

Why startup founders have mood swings (and why they may have uses)

(This post was collaboratively written together with Duncan Sabien.)

Startup founders stereotypically experience some pretty serious mood swings. One day, their product seems destined to be bigger than Google, and the next, it's a mess of incoherent, unrealistic nonsense that no one in their right mind would ever pay a dime for. Many of them spend half of their time full of drive and enthusiasm, and the other half crippled by self-doubt, despair, and guilt. Often this rollercoaster ride goes on for years before the company either finds its feet or goes under.



Well, sure, you might say. *Running a startup is stressful. Stress comes with mood swings.*

But that's not really an explanation—it's like saying stuff falls when you let it go. There's something about the "launching a startup" situation that induces these kinds of mood swings in *many* people, including plenty who would otherwise be entirely stable.

Moving parts

There are plenty of stressful situations which *don't* cause this sort of high-magnitude cycling. Waiting on blood test results from a cancer scare, or applying to prestigious colleges or companies, or working as a *subordinate* startup employee[1]—these are all demanding, emotionally draining circumstances, but they don't induce the same sort of cycling.

Contrast those situations with these:

- People in the early stages of their first serious romantic relationship (especially those who really really really want it to work, but lack a clear model of *how*)
- People who are deeply serious about personally making a difference in global poverty, space exploration, existential risk, or other world-scale issues
- People who are struggling to write their first novel, make their first movie, paint their first masterpiece, or win their first gold medal

... all of which have been known to push people into the same kind of oscillation we see in startup founders. As far as we can tell, there are two factors common to all of these cases: they're situations in which people must work really, really hard to have any hope of success, and they're *also* situations in which people aren't at all sure what *kinds* of work will get them there. It's not a response to high pressure **or** uncertainty, it's a response to high pressure **and** uncertainty, where uncertainty means not just being unsure about which path to take, but also about whether the paths (and the destination!) are even real.

Subconscious intentionality

It's easy to point to the value in euphoria and optimism. You get lots of code written, recruit lots of funding and talent, write a perfect draft—it's the part of the cycle where you're drawn to working seventy hour weeks, checking off each and every item from your to-do list. But the "down" parts often feel like they're pointless at best, and dangerous or counterproductive at worst. Most of the online commentary on this phenomenon treats them as a negative; there's lots of talk about how to break the pattern, or how to "deal with" the extremes. In our own pasts, we found ourselves wondering why our brains couldn't just hang on to the momentum—why they insisted on taking us through stupid detours of despair or shame before returning us back to apparent "forward motion".

Interesting things happened when we began treating that question as non-rhetorical. What, we wondered, might those down cycles be aimed at? If we were to *assume* that they were, on some level, useful/intentional/constructive, then what sorts of instrumental behavioral patterns were they nudging us toward?

Imagine breaking the two parts of the cycle into two separate people, each of whom is advising you as you move a startup forward. One of them is the voice of confidence—

the classic "inside view"—with a clear sense of the possibilities and lots of enthusiasm for next actions. The other is the voice of pessimism, with a range of doubts and concerns from the object level all the way up to the fundamental assumptions of your whole project.

A naive manager might assume that one of these advisors is better than the other. Certainly one of them is more “pleasant” to have around. But it’s almost trivially obvious that you’ll build better models and more robust plans if you keep both of them involved. It’s not just about balancing positive and negative outlooks, or constructive and destructive impulses (although that’s a part of it). It’s about the process of building an accurate map in unexplored territory, which often requires a kind of leapfrogging of provisional model-building and critical model-tearing-down. You start with a bad but workable model, collide it with reality, and then refine the model in response before colliding it again. Both halves of the process are critical—a good entrepreneur needs both the ability to follow a line of thinking to its conclusion *and* the ability to throw the line away and start from scratch, and would be equally crippled by a lack of either.

Why despair?

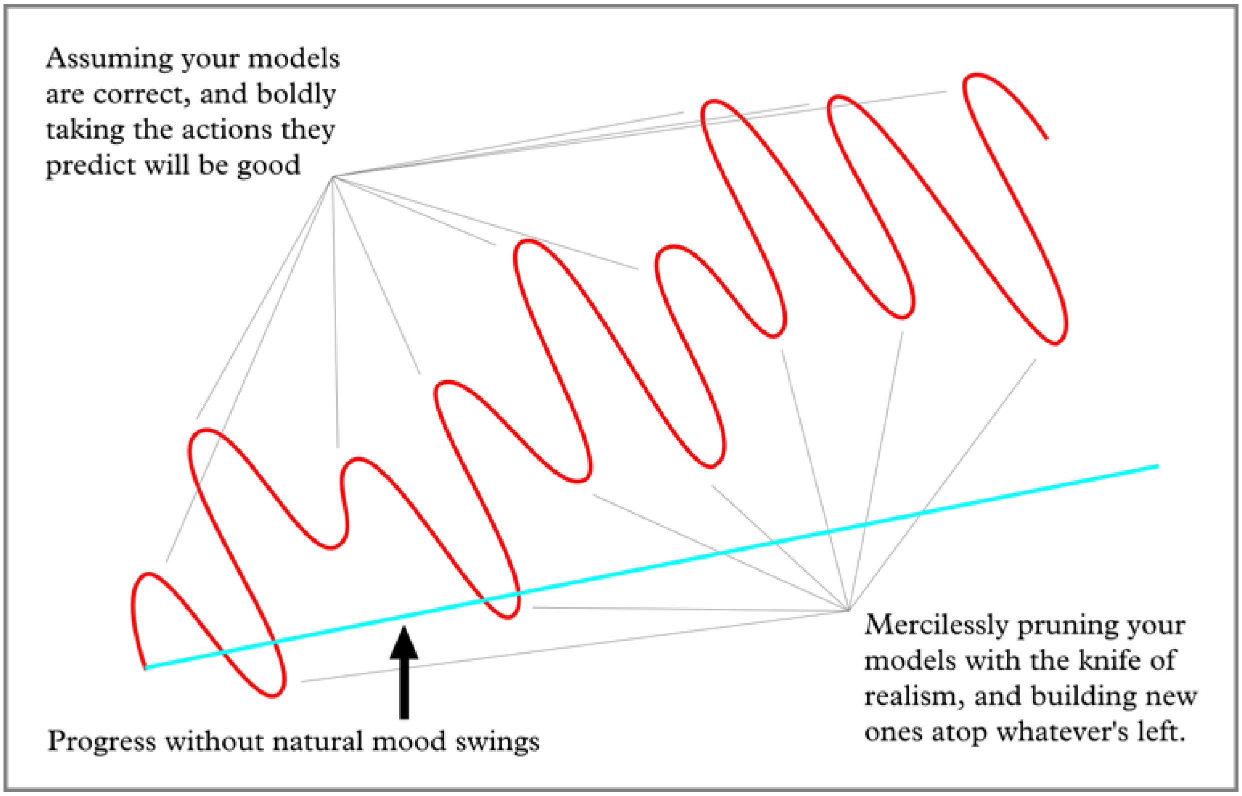
Startup founders, aspiring filmmakers, and would-be worldsavers “have to” be enthusiastic and positive. They have to *sell* their models—to skeptical investors, to potential hires, and oddly often to themselves—and many people struggle to hold both determined confidence and fundamental uncertainty in their minds at the same time. Instead, their brains use mood to encourage them to alternate. The euphoric half of the cycle is where assumptions are taken to their conclusion—where you collide your models with reality by building code, attempting to enroll customers, and so on. And the despair half is where those assumptions are challenged and questioned—where all of the fears and doubts and worrying bits of evidence that you’ve been holding off on looking at are allowed to come forward.

Interestingly, the real power of the down cycle seems to be less that it allows thoughts like “maybe my startup will never work,” and more that, *because* it allows such thoughts, it *also* allows “maybe my understanding of the sales landscape is wrong,” or “maybe the product manager we’ve just spent three months getting up to speed is actually a terrible fit for the team.” Despair can be a key that unlocks whole swaths of the territory. When you’re “up,” your current strategy is often weirdly entangled with your overall sense of resolve and commitment—we sometimes have a hard time critically and objectively evaluating parts C, D, and J because flaws in C, D, and J would threaten the whole edifice. But when you’re “down,” the obviousness of your impending doom means that you can look critically at your past assumptions without having to defend anything.[2]

Moving forward (or backward, as the case may be)

Our recommendation: the next time you're working on something large, important, and very, very uncertain, *don't resist the occasional slide into despair*. In fact, don't even think of it as a "slide"—in our experience, it stops feeling like a slide as soon as you stop resisting it. Instead, recognize it for what it is—a sign that important evidence has been building up in your buffer, unacknowledged, and that it's time now to integrate it into your plans.

Understanding the subconscious intentionality behind this sort of mood swing doesn't make it any less *painful*, but it can make it far more *pleasant*—easier to endure and easier to mine for value. Think of it as an efficient and useful process for building accurate models under stressful uncertainty, and trust yourself to hold together in both states, resisting neither.


Assuming your models are correct, and boldly taking the actions they predict will be good
Mercilessly pruning your models with the knife of realism, and building new ones atop whatever's left.
Progress without natural mood swings

There's a version of you that is good at moving forward—that has the energy to pour seventy hours a week into your current best guesses, stretching the parts of your model that are correct as far as they can go. And there's a version of you that is good at dealing with the darker side of reality—that is *actually willing* to consider the possibility that it's all garbage, instead of just paying lip service to the idea. Embrace each in turn, and the result can be periods of high productivity followed by periods of extreme flexibility, a combined strategy whose average utility often seems to outweigh the utility of an average strategy.

[1] The example of subordinate startup employees seems perhaps particularly revealing, since they share the same uncertainty and long hours as their bosses, but are not expected to come up with the same kinds of strategic solutions.

[2] Alternating between despair and forward motion may also be a good strategy given human working memory limitations. :)

Obvious advice

This is a linkpost for https://mindingourway.com/obvious-advice/

This is a common scene at the MIRI offices: I have a decision to make, like what sort of winter fundraiser to run. Before making any choices, I take a few minutes to write down all the obvious things to do before making the decision: spend five minutes brainstorming options before weighting any pros or cons; talk to people who have run different types of fundraisers in similar situations; and so on. I can usually generate a handful of obvious things to do before making my decision. I write those things down, and then I describe my decision to one of my advisors and see if they have any advice. They say "only the obvious," and then rattle off five more obvious things I hadn't thought of, all of them useful.

Sometimes, I wonder how successful a person would be if they just did all the obvious things in pursuit of their goals.

So with that in mind, allow me to offer some quite obvious pieces of advice, which have proven very useful for me:

Before carrying out any plan, *actually do the obvious things.*

When you're about to make a big decision, pause, and ask yourself what obvious things a reasonable person would do before making this sort of decision. Would they spend a full five minutes (by the clock) brainstorming alternative options before settling on a decision? Would they consult with friends and advisors? Would they do some particular type of research?

Then, *actually do the obvious things.*

A corollary to this advice is to also occasionally consider *not doing things the wrong way.* Imagine someone who's recently failed at an endeavor that was important to them. They're fraught with despair, and you attempt to console them by saying "well, at least you learned something." They snap back, "yeah, I learned never to try hard things ever again!"

This may be just an emotional outburst, yes, but if they act upon this outburst — and withdraw, and become less curious, and become more bitter — then they are in solid need of the above corollary. In fact, the middle of an emotional outburst is one of the *best times* to use the corollary. I have often myself found it useful, mid-hasty-decision, to pause, reflect, and ask myself "wait… is this a *terrible plan?*"

(And then, if the answer is yes, I don't carry out the plan — a crucial step.)

Both pieces of advice above — "do the obvious preparation", and "don't execute bad plans" — each get a lot more useful as you expand your notions of "obvious preparation" and "bad plan". In fact, quite a bit of rationalist-style advice is about expanding your notion of "obvious thing" and "bad plan." Thus, this advice gets much more helpful if you make sure to do the obvious things.

(Not all the rationalist advice is of this form, of course; many of the most important rationalist skills are cognitive operations that happen in five seconds or less. One example of a five-second-level skill is the skill of encountering a new problem and *reflexively starting to list obvious perparations* or noticing an emotional outburst and *reflexively taking a step back and checking whether your current plan is terrible.* More often than not, one of the goals of these blog posts is to install a five-second cognitive reflex of *deciding to apply a tool* by describing the tool itself. But I digress.)

For example, the cognitive reflex of "enumerate obvious preparations" becomes much more useful once you have concepts like "brainstorm options before weighing pros and cons" and "set a five minute timer and actually think about the problem for the whole five minutes" and "consider the opportunity costs." And the cognitive reflex of "check whether your current plan is terrible" becomes more useful as you add concepts like "rationalizing" and "blindly acting out a social role" and so on.

So this week's advice is obvious advice, but useful nonetheless: find a way to gain a reflex to actually do all the obvious preparation, before undertaking a new task or making a big decision.

It's surprising how often the advice that I give people who come to me asking for advice cashes out to some form of "well, have you considered doing the obvious thing?"

For example, when someone comes to me and says "help, I have a talk I have to give and I'm going to be terribly nervous and I dread it, what do I do?" it's often surprisingly helpful for me to ask, "well, what sort of things would make you less nervous?" Or someone comes to me and says "I find myself just playing video games all day, how do I stop myself?", I first ask, "have you considered what sorts of things you'd rather do besides play video games all day?"

In many cases, the obvious prompts aren't sufficient. But in a surprising number of cases, *they are.* I still often find this advice useful myself: when my attention slips, I am often helped by someone just asking me to consider the obvious — "what would make the task less dreadful?" or "have you thought for five minutes about alternatives?" or "have you considered delegating this?" and so on.

Much of my advice for how to manage guilt was generated by this very process, by me imagining feeling guilty, and then imagining which obvious things I'd try to do to engage with the feeling. I would ask myself questions like "what is the cause of this feeling?" and "how is it being useful to me?" and "is there a better way I can achieve those goals?" and I would spend time listening to myself and brainstorming options, because those are all the obvious ways to address the problem. Many of my early posts on guilt were a product of articulating that reflexive process. The types of obvious advice that I would generate — such as asking "what is the cause and use of this feeling?" — might be very different from the obvious advice that you would generate, and that's fine. The trick is to apply the obvious advice first.

Or imagine you have the problem of finding it difficult to use the "do the obvious" technique. Maybe you've been struggling to remember to consider the obvious whenever you encounter a hard decision. Instead of asking for advice, consider generating a list like the following, first:

- Spend five minutes generating examples of decisions you made in the past where it would have been helpful to do the obvious things first. Then spend five minutes examining those and looking for patterns.

- Close your eyes and visualize yourself facing a new decision in as much concrete detail as you can, and practice thinking "oh wait, let me list the obvious things before proceeding."
- Train yourself to notice decision-points better by buying a tally counter and tracking decisions and giving yourself positive reinforcement every time you do.

Or imagine that you have tried to do all the obvious things, and you find that you're going into "enumerate the obvious" mode even for the most trivial tasks, and it's making your trivial tasks take way too long and the whole thing seems pretty foolish. Then, before complaining, consider trying the corollary, and consider whether applying "try the obvious" far too often is in fact a terrible plan.

Your list of obvious things will very likely look very different from my own — my friends and advisors *still* generate obvious-in-retrospect ideas that I myself was incapable of generating, even after spending a few minutes generating the sort of ideas I expected them to generate. Collecting tips and techniques from other people in your environment is a great way to expand your "obvious things" repertoire, and asking for advice from friends will likely continue to generate new obvious things for quite some time. It's OK for your lists to be very different; the trick is to do *some* of the obvious preparation before making a hard decision. It can often make a difference.

The important thing, here, is to find a way to actually start doing the obvious things. This is the skill that's like footwork for a rationalist: remembering to actually do the obvious preparation is easy to learn and difficult to master; it's a skill to drill when you have spare mental energy in hopes that it comes naturally and easily whenever the going gets tough and the stakes get high.

I continue to wonder how powerful a person could become, if they simply managed to do all the obvious things in pursuit of their goals.