Best of LessWrong: January 2016

1. The correct response to uncertainty is *not* half-speed
2. [moderator action] The_Lion and The_Lion2 are banned
3. Why CFAR's Mission?
4. Anxiety and Rationality
5. Confidence all the way up
6. Desperation
7. The art of response

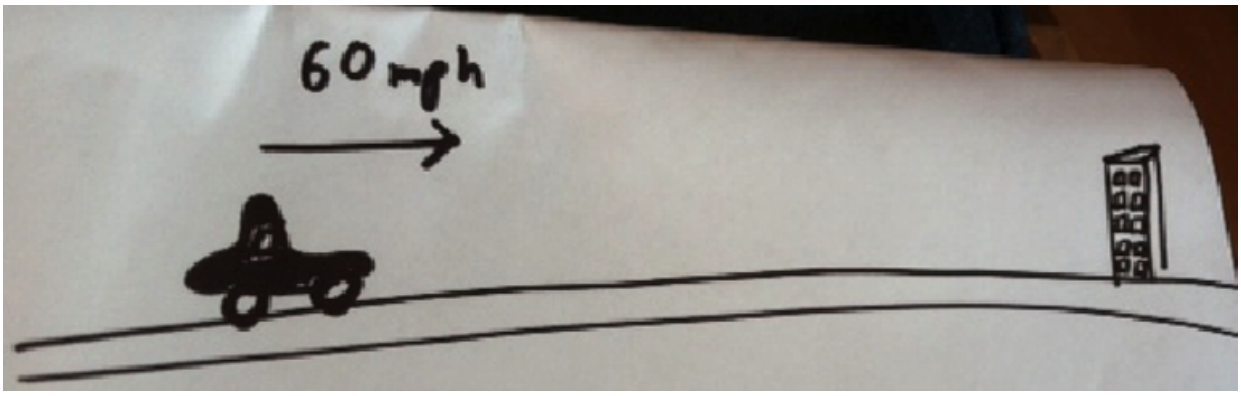
Best of LessWrong: January 2016

1. The correct response to uncertainty is *not* half-speed
2. [moderator action] The_Lion and The_Lion2 are banned
3. Why CFAR's Mission?
4. Anxiety and Rationality
5. Confidence all the way up
6. Desperation
7. The art of response

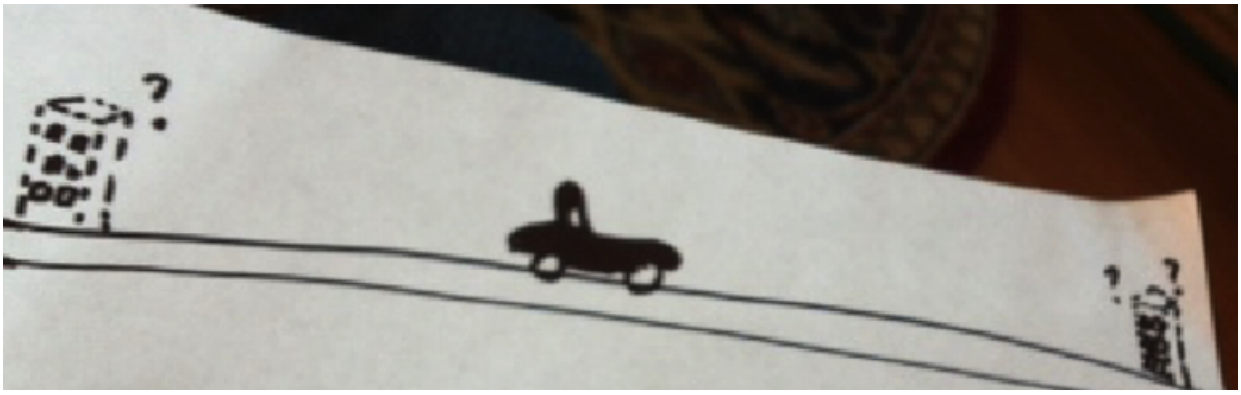
The correct response to uncertainty is *not* half-speed

Related to: Half-assing it with everything you've got; Wasted motion; Say it Loud.

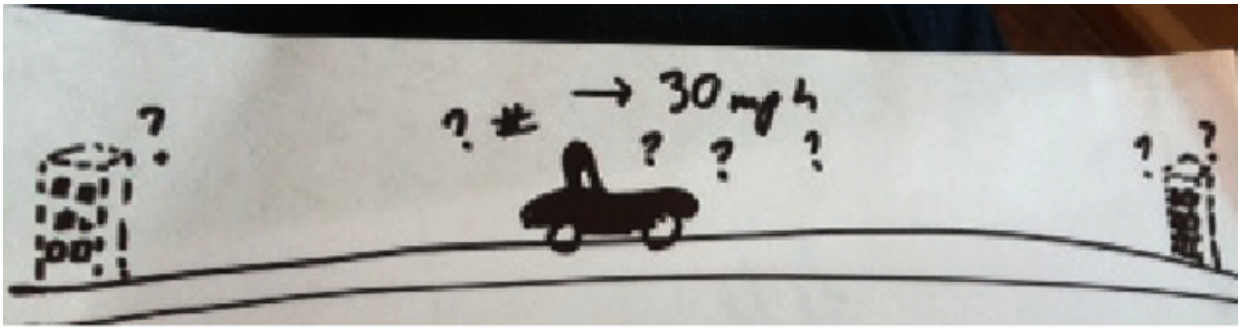
Once upon a time (true story), I was on my way to a hotel in a new city. I knew the hotel was many miles down this long, branchless road. So I drove for a long while.



After a while, I began to worry I had passed the hotel.



So, instead of proceeding at 60 miles per hour the way I had been, I continued in the same direction for several more minutes at 30 miles per hour, wondering if I should keep going or turn around.



After a while, I realized: I was being silly! If the hotel was ahead of me, I'd get there fastest if I kept going 60mph. And if the hotel was behind me, I'd get there fastest by heading at 60 miles per hour in the other direction. And if I *wasn't* going to turn around yet -- if my best bet given the uncertainty was to check *N* more miles of highway first, *before* I turned around -- then, again, I'd get there fastest by choosing a value of *N*, speeding along at 60 miles per hour until my odometer said I'd gone *N* miles, and *then* turning around and heading at 60 miles per hour in the opposite direction.

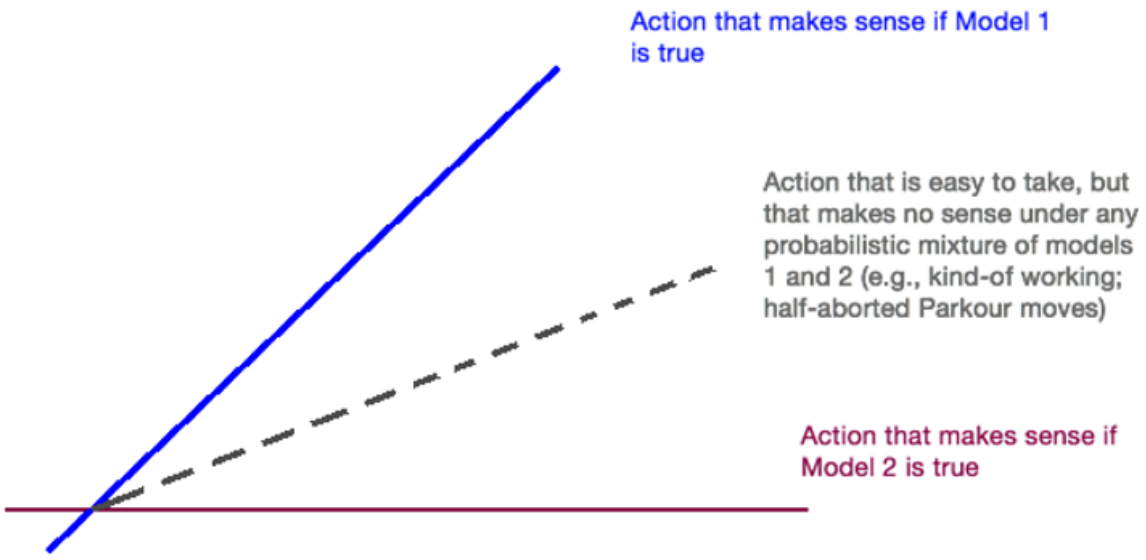
Either way, fullspeed was best. My mind had been naively averaging two courses of action -- the thought was something like: "maybe I should go forward, and maybe I should go backward. So, since I'm uncertain, I should go forward at half-speed!" But averages don't *actually* work that way.[1]

Following this, I started noticing lots of hotels in my life (and, perhaps less tactfully, in my friends' lives). For example:

- I wasn't sure if I was a good enough writer to write a given doc myself, or if I should try to outsource it. So, I sat there kind-of-writing it while also fretting about whether the task was correct.
    - (Solution: Take a minute out to think through heuristics. Then, either: (1) write the post at full speed; or (2) try to outsource it; or (3) write full force *for some fixed time period*, and then pause and evaluate.)
- I wasn't sure (back in early 2012) that CFAR was worthwhile. So, I kind-of worked on it.
- An old friend came to my door unexpectedly, and I was tempted to hang out with her, but I also thought I should finish my work. So I kind-of hung out with her while feeling bad and distracted about my work.
- A friend of mine, when teaching me math, seems to mumble *specifically those words that he doesn't expect me to understand* (in a sort of compromise between saying them and not saying them)...
- Duncan reports that novice Parkour students are unable to safely undertake certain sorts of jumps, because they risk aborting the move mid-stream, after the *actual* last safe stopping point (apparently kind-of-attempting these jumps is more dangerous than either attempting, or not attempting the jumps)
- It is said that start-up founders need to be irrationally certain that their startup will succeed, lest they be unable to do more than kind-of work on it...

That is, it seems to me that often there are two different actions that would make sense under two different models, and we are uncertain which model is true... and so

we find ourselves taking an intermediate of half-speed action... even when that action makes no sense under any probabilistic mixture of the two models.



You might try looking out for such examples in your life.

---

[1] Edited to add: The hotel example has received much nitpicking in the comments. But: (A) the actual example was legit, I think. Yes, stopping to think has some legitimacy, but driving slowly for a long time because uncertain does not optimize for thinking. Similarly, it may make sense to drive slowly to stare at the buildings in some contexts... but I was on a very long empty country road, with no buildings anywhere (true historical fact), and also I was not squinting carefully at the scenery. The thing I needed to do was to execute an efficient search pattern, with a threshold for a future time at which to switch from full-speed in some direction to full-speed in the other. Also: (B) consider some of the other examples; "kind of working", "kind of hanging out with my friend", etc. seem to be common behaviors that are mostly not all that useful in the usual case.

[moderator action] The_Lion and The_Lion2 are banned

Accounts "The_Lion" and "The_Lion2" are banned now. Here is some background, mostly for the users who weren't here two years ago:

User "Eugine_Nier" was banned for retributive downvoting in July 2014. He keeps returning to the website using new accounts, such as "Azathoth123", "Voiceofra", "The_Lion", and he keeps repeating the behavior that got him banned originally.

The original ban was **permanent**. It *will* be enforced on all future known accounts of Eugine. (At random moments, because moderators sometimes feel too tired to play whack-a-mole.) This decision is *not* open to discussion.

Please note that the moderators of LW are the opposite of trigger-happy. Not counting spam, there is on average *less than one account per year* banned. I am writing this explicitly, to avoid possible misunderstanding among the new users. Just because you have read about someone being banned, it doesn't mean that *you* are now at risk.

Most of the time, LW discourse is regulated by the community voting on articles and comments. Stupid or offensive comments get downvoted; you lose some karma, then everyone moves on. In rare cases, moderators may remove specific content that goes against the rules. The account ban is only used in the extreme cases (plus for obvious spam accounts). Specifically, on LW people *don't* get banned for merely not understanding something or disagreeing with someone.

What does "retributive downvoting" mean? Imagine that in a discussion you write a comment that someone disagrees with. Then in a few hours you will find that your karma has dropped by hundreds of points, because someone went through your entire comment history and downvoted all comments you ever wrote on LW; most of them completely unrelated to the debate that "triggered" the downvoter.

Such behavior is damaging to the debate and the community. Unlike downvoting a specific comment, this kind of mass downvoting isn't used to correct a *faux pas*, but to drive a *person* away from the website. It has especially strong impact on new users, who don't know what is going on, so they may mistake it for a reaction of the whole community. But even in experienced users it creates an "ugh field" around certain topics known to invoke the reaction. Thus a single user has achieved disproportional control over the content and the user base of the website. This is not desired, and will be punished by the site owners and the moderators.

To avoid *rules lawyering*, there is no exact definition of how much downvoting breaks the rules. The rule of thumb is that you should upvote or downvote each comment based on the value of that specific comment. You shouldn't vote on the comments regardless of their content merely because they were written by a specific user.

Why CFAR's Mission?

Related to:

- Our 2015 Fundraiser; Why CFAR? The view from 2015; Critch's "The 2015 x-risk ecosystem".

Briefly put, CFAR's mission is to improve the sanity/thinking skill of those who are most likely to actually usefully impact the world.

I'd like to explain what this mission means to me, and why I think a high-quality effort of this sort is essential, possible, and urgent.

I used a Q&A format (with imaginary Q's) to keep things readable; I would also be very glad to Skype 1-on-1 if you'd like something about CFAR to make sense, as would Pete Michaud. You can schedule a conversation automatically with me or Pete.

Q: Why not focus exclusively on spreading altruism? Or else on "raising awareness" for some particular known cause?

Briefly put: because historical roads to hell have been powered in part by good intentions; because the contemporary world seems bottlenecked by its ability to *figure out what to do and how to do it* (i.e. by ideas/creativity/capacity) more than by folks' willingness to sacrifice; and because rationality skill and epistemic hygiene seem like skills that may distinguish *actually useful* ideas from ineffective or harmful ones in a way that "good intentions" cannot.

Q: Even given the above -- why focus extra on sanity, or *true* beliefs? Why not focus instead on, say, competence/usefulness as the key determinant of how much do-gooding impact a motivated person can have? (Also, have you ever met a Less Wronger? I hear they are annoying and have lots of problems with "akrasia", even while priding themselves on their high "epistemic" skills; and I know lots of people who seem "less rational" than Less Wrongers on some axes who would nevertheless be more useful in many jobs; is this "epistemic rationality" thingy *actually* the thing we need for this world-impact thingy?...)

This is an interesting one, IMO.

Basically, it seems to me that epistemic rationality, and skills for forming accurate explicit world-models, become more useful the more ambitious and confusing a problem one is tackling.

For example:

> If I have one floor to sweep, it would be best to hire a person who has pre-existing skill at sweeping floors.
>
> If I have 250 floors to sweep, it would be best to have someone energetic and perceptive, who will stick to the task, notice whether they are succeeding, and

> improve their efficiency over time. An "all round competent human being", maybe.
>
> If I have 10^25 floors to sweep, it would... be rather difficult to win at all, actually. But if I *can* win, it probably *isn't* by harnessing my pre-existing skill at floor-sweeping, nor even (I'd claim) my pre-existing skill at "general human competence". It's probably by using the foundations of science and/or politics to (somehow) create some totally crazy method of getting the floors swept (a process that would probably require actually accurate beliefs, and thus epistemic rationality).

The world's most important problems look to me more like that third example. And, again, it seems to me that to solve problems of *that* sort -- to iterate through many wrong guesses and somehow piece together an accurate model until one finds a workable pathway for doing what originally looked impossible -- without getting stuck in dead ends or wrong turns or inborn or societal prejudices -- it is damn helpful to have something like epistemic rationality. (Competence is pretty darn helpful too -- it's good to e.g. be able to go out there and get data; to be able to form networking relations with folks who already know things; etc. -- but epistemic rationality is necessary in a more fundamental way.)

For the sake of concreteness, I will claim that AI-related existential risk is among humanity's most important problems, and that it is damn confusing, damn hard, and really really needs something like epistemic rationality and not just something like altruism and competence if one is to impact it *positively*, rather than just, say, randomly impacting it. I'd be glad to discuss in the comments.

Q: Why suppose "sanity skill" *can* be increased?

Let's start with an easier question: why suppose thinking skills (of any sort) can be increased?

The answer to that one is easy: Because we see it done all the time.

The math student who arrives at college and does math for the first time with others is absorbing a kind of thinking skill; thus mathematicians discuss a person's "mathematical maturity", as a property distinct from (although related to) their learning of this and that math theorem.

Similarly, the coder who hacks her way through a bunch of software projects and learns several programming languages will have a much easier time learning her 8th language than she did her first; basically because, somewhere along the line, she learned to "think like a computer scientist"...

The claim that "sanity skill" is a type of thinking skill and that it can be increased is somewhat less obvious. I am personally convinced that the LW Sequences / AI to Zombies gave me something, and gave something similar to others I know, and that hanging out in person with Eliezer Yudkowsky, Michael Vassar, Carl Shulman, Nick Bostrom, and others gave me more of that same thing; a "same thing" that included e.g. actually trying to figure it out, making beliefs pay rent in anticipated experience; using arithmetic to entangle different pieces of my beliefs; and so on.

I similarly have the strong impression that e.g. Feynman's and Munger's popular writings often pass on pieces of this same thing; that the convergence between the LW Sequences and Tetlock's Superforecasting training is non-coincidental; that the

convergence between CFAR's workshop contents and a typical MBA program's contents is non-coincidental (though we were unaware of it when creating our initial draft); and more generally that there are many types of thinking skill that are routinely learned/taught and that non-trivially aid the process of coming to accurate beliefs in tricky domains. I update toward this partly from the above convergences; from the fact that Tetlock's training seems to work; from the fact that e.g. Feynman and Munger (and for that matter Thiel, Ray Dalio, Francis Bacon, and a number of others) were shockingly conventionally successful and advocated similar things; and from the fact that there is quite a bit of "sanity" advice that is obviously correct once stated, but that we don't automatically do (advice like "bother to look at the data; and try to update if the data doesn't match your predictions").

So, yes, I suspect that there is some portion of sanity that can sometimes be learned and taught. And I suspect this portion can be increased further with work.

Q. Even if you can train skills: Why go through all the trouble and complications of trying to do this, rather than trying to find and recruit people who *already* have the skills?

The short version: because there don't seem to be enough people out there with the relevant skills (yet), and because it does seem to be possible to help people increase their skills with training.

How can I tell there aren't enough people out there, instead of supposing that we haven't yet figured out how to find and recruit them?

Basically, because it seems to me that if people had really huge amounts of epistemic rationality + competence + caring, they would already be impacting these problems. Their huge amounts of epistemic rationality and competence would allow them to find a path to high impact; and their caring would compel them to do it.

I realize this is a somewhat odd thing to say, and may not seem true to most readers. It didn't used to seem true to me. I myself found the idea of existential risk partly through luck, and so, being unable to picture thinkers who had skills I lacked, I felt that anyone would need luck to find out about existential risk. Recruiters, I figured, could fix that luck-gap by finding people and telling them about the risks.

However, there are folks who can reason their own way to the concept. And there are folks who, having noticed how much these issues matter, can locate novel paths that may plausibly impact human outcomes. And even if there weren't folks of this sort, there would be hypothetical people of this sort; much of what is luck at some skill-levels can in principle be made into skill. So, finding that only a limited number of people are working effectively on these problems (provided one is right about that assessment) does offer an upper bound on how many people today can exceed a certain combined level of epistemic rationality, competence, and caring about the world.

And of course, we should *also* work to recruit competence and epistemic skill and caring wherever we can find them. (To help such folk find their highest impact path, whatever that turns out to be.) We just shouldn't expect such recruiting to be enough, and shouldn't stop there. Especially given the stakes. And given the things that may happen if e.g. political competence, "raised awareness", and recruited good intentions are allowed outpace insight and good epistemic norms on e.g. AI-related existential

risk, or other similarly confusing issues. Andrew Critch makes this point well with respect to AI safety in a recent post; I agree with basically everything he says there.

Q. Why do you sometimes talk of CFAR's mission as "improving thinking skill", and other times talk of improving all of epistemic rationality, competence, and do-gooding? Which one are you after?

The main goal is thinking skill. Specifically, thinking skill among those most likely to successfully use it to positively impact the world.

Competence and caring are relevant secondary goals: some of us have a conjecture that deep epistemic rationality can be useful for creating competence and caring, and of course competence and caring about the world are also directly useful for impacting the world's problems. But CFAR wants to increase competence and caring *via* teaching relevant pieces of thinking skill, and not via special-case hacks. For example, we want to help people stay tuned into what they care about even when this is painful, and to help people notice their aversions and sort through which of their aversions are and aren't based in accurate implicit models. We do not want to use random emotional appeals to boost specific cause areas, nor to use other special-case hacks that happen to boost efficacy in a manner opaque to participants.

Why focus primarily on thinking skill? Partly so we can have focus enough as an organization so as to actually do anything at all. (Organizations that try to accomplish several things at once risk accomplishing none -- and "epistemic rationality" is more of a single thing.) Partly so our workshop participants and other learners can similarly have focus as learners. And partly because, as discussed above, it is very very hard to intervene in global affairs in such a way as to *actually have **positive** outcomes,* and not merely outcomes one pretends will be positive; and focusing on actual thinking skill seems like a better bet for problems as confusing as e.g. existential risk.

Why include competence and caring at all, then? Because high-performing humans make use of large portions of their minds (I think), and if we focus only on "accurate beliefs" in a narrow sense (e.g., doing analogs of Tetlocks forecasting training and nothing else), we are apt to generate "straw lesswrongers" whose "rationality" applies mainly to their explicit beliefs... people who can nitpick incorrect statements and can in this way attempt accurate verbal statements, but who are not creatively generative, do not have the iterated energy/competence/rapid iteration required to launch a startup, and cannot run good fast realtime social skills. We aim to do better. And we suspect that working to hit competence and caring *via* what one might call "deep epistemic rationality" is a route in.

Q. Can a small organization realistically do all that without losing Pomodoro virtue? (By "Pomodoro virtue", I mean the ability to focus on one thing at a time and so to actually make progress, instead of losing oneself amidst the distraction of 20 goals.)

We think so, and we think the new core/labs division within CFAR will help. Basically, Core will be working to scale up the workshops and related infrastructure, which should give a nice trackable set of numbers to optimize -- numbers that, if grown, will enable better financial health for CFAR and will also enable a much larger set of people to train in rationality.

Labs will be focusing on impacting smaller numbers of people who are poised to impact existential risk (mainly), and on seeing whether our "impacts" on these folk do

in fact seem to help with their impact on the world.

We will continue to work together on many projects and to trade ideas frequently, but I suspect that this separation into two goals will give more "pomorodo virtue" to the whole organization.

Q. What is CFAR's relationship to existential risk? And what should it be?

CFAR's mission is to improve the sanity/thinking skill of those who are most likely to actually usefully impact the world -- via whatever cause areas may be most important. Many of us suspect that AI-related existential risk is an area with huge potential for useful impact; and so we are focusing partly on helping meet talent gaps in that field. This focus also gives us more "pomodoro virtue" -- it is easier to track whether e.g. the MIRI Summer Fellows Program helped boost research on AI safety, than it is to track whether a workshop had "good impacts on the world" in some more general sense.

It is important to us that the focus remain on "high impact pathways, whatever those turn out to be", that we do not propagandize for particular pre-set answers (rather, that we assist folks in thinking things through in an unhindered way), and that we work toward a kind of thinking skill that may let people better assess what paths are *actually* high impact for having positive effects in the world, and to overcome flaws in our current thinking.

Q. Should I do “Earning to Give”? Also: I heard that there are big funders around now and so “earning to give” is no longer a sensible thing for most people to do; is that true? And what does all this have to do with CFAR?

IMO, earning to give remains a pathway for creating very large positive impacts on the world.

For example, CFAR, MIRI, and many of the organizations under the CEA umbrella seem to me to be both high potential impact, and substantially funding limited right now, such that further donation is likely to cause a substantial increase in how much good these organizations accomplish.

This is an interesting claim to make in a world where e.g. Good Ventures, Elon Musk, and others are already putting very large sums of money into making the world better; if you expect to make a large difference via individual donation, you must implicitly expect that you can pick donation targets for your own smallish sum better than they could, at least at the margin. (One factor that makes this more plausible is that you can afford to put in far more time/attention/thought per dollar than they can.)

Alternately stated: the world seems to me to be short, not so much on money, as on understanding of what to do with money; and an "Earning to Give" career seems to me to potentially make sense insofar as it is a decision to *really actually try to figure out* how to get humanity to a win-state (or how to otherwise do whatever it is that is actually worthwhile), especially insofar as you seem to see low-hanging fruit for high-impact donation. Arguing with varied others who have thought about it is perhaps the fastest route toward a better-developed inside view, together with setting a 5 minute timer and attempting to solve it yourself. "Earning to give"ers, IMO, contribute in proportion to the amount of non-standard epistemic skill they develop, and not just in proportion to their giving amount; and, much as with Giving Games, knowing that you matter today can provide a reason to develop a worldview for real.

This has to do with CFAR both because I expect us to accomplish far more good if we have money to do it, and because folks *actually trying to figure out how to impact the world* are building the kind of thinking skill CFAR is all about.

Anxiety and Rationality

Recently, someone on the Facebook page asked if anyone had used rationality to target anxieties. I have, so I thought I'd share my LessWrong-inspired strategies. This is my first post, so feedback and formatting help are welcome.

First things first: the techniques developed by this community are not a panacea for mental illness. They *are* way more effective than chance and other tactics at reducing normal bias, and I think many mental illnesses are simply cognitive biases that are extreme enough to get noticed. In other words, getting a probability question about cancer systematically wrong does not disrupt my life enough to make the error obvious. When I believe (irrationally) that I will get fired because I asked for help at work, my life is disrupted. I become non-functional, and the error is clear.

Second: the best way to attack anxiety is to do the things that make your anxieties go away. That might seem too obvious to state, but I've definitely been caught in an "analysis loop," where I stay up all night reading self-help guides only to find myself non-functional in the morning because I didn't sleep. If you find that attacking an anxiety with Bayesian updating is like chopping down the Washington monument with a spoon, but getting a full night's sleep makes the monument disappear completely, consider the sleep. Likewise for techniques that have little to no scientific evidence, but are a good placebo. A placebo effect is still an effect.

Finally, like all advice, this comes with Implicit Step Zero: "Have enough executive function to give this a try." If you find yourself in an analysis loop, you may not yet have enough executive function to try any of the advice you read. The advice for functioning *better* is not always identical to the advice for functioning *at all.* If there's interest in an "improving your executive function" post, I'll write one eventually. It will be late, because my executive function is not impeccable.

Simple updating is my personal favorite for attacking *specific* anxieties. A general sense of impending doom is a very tricky target and does not respond well to reality. If you can narrow it down to a particular belief, however, you can amass evidence against it.

Returning to my example about work: I *alieved* that I would get fired if I asked for help or missed a day due to illness. The distinction between *believe* and *alieve* is an incredibly useful tool that I immediately integrated when I heard of it. Learning to make beliefs pay rent is much easier than making harmful aliefs go away. The tactics are similar: do experiments, make predictions, throw evidence at the situation until you get closer to reality. Update accordingly.

The first thing I do is identify the situation and why it's dysfunctional. The alief that I'll get fired for asking for help is not actually articulated when it manifests as an anxiety. Ask me in the middle of a panic attack, and I *still* won't articulate that I am afraid of getting fired. So I take the anxiety all the way through to its implication. The algorithm is something like this:

1. Notice sense of doom
2. Notice my avoidance behaviors (not opening my email, walking away from my desk)
3. Ask "What am I afraid of?"
4. Answer (it's probably silly)

5. Ask “What do I think will happen?”
6. Make a prediction about what will happen (usually the prediction is implausible, which is why we want it to go away in the first place)

In the “asking for help” scenario, the answer to “what do I think will happen” is implausible. It’s extremely unlikely that I’ll get fired for it! This helps take the gravitas out of the anxiety, but it does not make it go away.* After (6), it’s usually easy to do an experiment. If I ask my coworkers for help, will I get fired? The only way to know is to try.

…That’s actually not true, of course. A sense of my environment, my coworkers, and my general competence at work should be enough. But if it was, we wouldn’t be here, would we?

So I perform the experiment. And I wait. When I receive a reply of any sort, even if it’s negative, I make a tick mark on a sheet of paper. I label it “didn’t get fired.” Because again, even if it’s negative, I didn’t get fired.

This takes a lot of tick marks. Cutting down the Washington monument with a spoon, remember?

The tick marks don’t have to be physical. I prefer it, because it makes the “updating” process visual. I’ve tried making a mental note and it’s not nearly as effective. Play around with it, though. If you’re anything like me, you have a lot of anxieties to experiment with.

Usually, the anxiety starts to dissipate after obtaining several tick marks. Ideally, one iteration of experiments should solve the problem. But we aren’t ideal; we’re mentally ill. Depending on the severity of the anxiety, you may need someone to remind you that doom will not occur. I occasionally panic when I have to return to work after taking a sick day. I ask my husband to remind me that I won’t get fired. I ask him to remind me that he’ll still love me if I do get fired. If this sounds childish, it’s because it is. Again: we’re mentally ill. Even if you aren’t, however, assigning value judgements to essentially harmless coping mechanisms does not make sense. Childish-but-helpful is much better than mature-and-harmful, if you have to choose.

I still have tiny ugh fields around my anxiety triggers. They don’t really go away. It’s more like learning not to hit someone you’re angry at. You notice the impulse, accept it, and move on. Hopefully, your harmful alief starves to death.

If you perform your experiment and doom *does* occur, it might not be you. If you can’t ask your boss for help, it might be your boss. If you disagree with your spouse and they scream at you for an hour, it might be your spouse. This isn’t an excuse to blame your problems on the world, but abusive situations can be sneaky. Ask some trusted friends for a sanity check, if you’re performing experiments and getting doom as a result. This is designed for situations where your alief is obviously silly. Where you *know* it’s silly, and need to throw evidence at your brain to internalize it. It’s fine to be afraid of genuinely scary things; if you really are in an abusive work environment, maybe you *shouldn’t* ask for help (and start looking for another job instead).

*using this technique for several months occasionally stops the anxiety immediately after step 6.

Confidence all the way up

This is a linkpost for https://mindingourway.com/confidence-all-the-way-up/

I apparently possess some sort of aura of competence. Some say I'm confident, others say I'm arrogant, others remark on how I seem very certain of myself (which I have been told both as compliment and critique).

I was surprised, at first, by these remarks from friends and family — from my perspective, I'm usually the first person in the conversation to express uncertainty in the form of probability estimates and error bars. I'm often quick to brainstorm alternative explanations of the data I use to support my claims. And, of course, I'm certain of nothing.

In fact, I had a conversation with a friend about this phenomenon once, which went something like this:

> ***Me:*** *Hey, have you noticed how everyone thinks I have an aura of confidence and certainty, sometimes arrogance? I don't know how to shake it, nor how it works. What's up with that?*
>
> ***Him:*** *Well, you always seem to have a solid grasp on every situation. When you're explaining things, you answer questions quickly, deftly, and with precision.*
>
> ***Me:*** *I don't think that's it, though. I'm rarely confident in the claims I'm making, and I tend to highlight that fact. Earlier, when we were talking with [other friend] about tools society can use to break monopolies, I was very explicit about where my uncertainty lies, and what assumptions my models relied upon, and where they might be flawed.*
>
> ***Him:*** *Yeah, but even then you were confident in what you were saying — maybe not confident in any particular claim you made, but confident in your overall analysis.*
>
> ***Me:*** *I don't think that's it either. I'll be the first to admit that the probabilities I put on my propositions are pulled out of thin air, and I'll also be the first to admit that my hypothesis space is decrepit and that I'd be able to find better models if I could think better. In fact, I'm aware of a bunch of flaws in the ways I think, and I dedicate a decent amount of effort to improving my own reasoning methods.*
>
> ***Him:*** *...*
>
> ***Me:*** *... I'm doing the thing right now, aren't I?*
>
> ***Him:*** *Yes, yes you are.*

There definitely is something of "confidence" to this pattern of speech and thinking, but it's not an empirical confidence. The confidence people notice in me isn't in the *content* of my claims, for I'm quick to couch my claims with probability estimates and error bars. Most of the confidence isn't in my analysis, either; I'm quick to note the ways my analyses could be flawed.

Some of the confidence *does* reside in the ways I reason; I do admit that I am much better equipped to answer questions of the form "but why are you so much more confident in your own reasoning than their reasoning, when they actually have more credentials?" than most. But even there, I can note plausible biases and judgement errors in my own reasoning processes with alacrity.

Why, then, do I come off as so confident? Why do I seem so self-assured while listing the ways I know my brain is flawed?

On reflection, I've concluded that (at least part of) the answer is something I call "confidence all the way up". Insofar as I'm uncertain of my content, I'm confident in my analysis — except, I'm not fully confident in my analysis. But insofar as I'm uncertain of my analysis, I'm confident in my reasoning procedures — except, I don't put faith there, either. But insofar as I'm uncertain of my reasoning procedures, I'm confident in my friends and failsafe mechanisms that will eventually force me to take notice and to update. Except, that's not quite right either — it's more like, every lack of confidence is covered by confidence one meta-level higher in the cognitive chain.

The result is something that reads socially as confidence regardless of how much empirical uncertainty I'm under.

Where does it bottom out? Well, insofar as my friends and failsafe mechanisms aren't sufficient to raise errors to my attention, I expect to reason poorly in an irredeemable fashion and then fail to achieve my goals. It bottoms out at the point where I say "yeah, if I'm *that far gone*, then I fail and die."

(And somehow, I'm able to say even this while maintaining my aura of self-assuredness and confidence).

I have encountered many people who seem paralyzed by their uncertainties. They hit a question (such as "what methods can a society use to break up monopolies?") and they are pretty sure that they won't be able to generate the *right* answers, and so they generate *no* answers.

And this may be a better failure mode than the failure mode of someone who has *too much* confidence and self-assuredness, who makes up a bunch of bad answers and then believes them with all their heart.

Someone with Confidence All The Way Up, though, can achieve the third alternative: generate a bunch of bad answers, understand why they're bad and where their limitations are, and use that information as best they can.

I have found this mindset to be very useful throughout my life. Confidence all the way up is what has me dive into the fray to try new things, while others stand on the sidelines bemoaning a high degree of uncertainty. It's part of the technique of treat recurring failures as data and training, rather than as a signal that it's time to feel guilty. It's part of the technique of knowing you're deeply limited without letting that interfere with your progress towards the goal. Of the top ten most competent people I've met in person (by my estimation), eight of them seem to have some variant of confidence all the way up running. If the mindset seems foreign to you, I suggest finding a way to practice it for a while.

Confidence all the way up is about working with what you have. It's about knowing your limitations. It's about knowing that you don't have perfect models of "what you have" nor "your limitations", and proceeding anyway, with an even stride.

It's about knowing that there are going to be curveballs, and trusting your ability to handle curveballs, but not all the time; and trusting your ability to get back up when you're knocked down by a curveball you couldn't handle, but not all the time; and coming to terms with the fact that you might be hurt so badly you can't get up.

Yes, we're limited. All humans are limited. There are important, decision-relevant facts that we don't know. Our reasoning processes run on compromised hardware. But the correct response to uncertainty is not to proceed at half speed!

No matter how hard you try to justify your beliefs, if you're being honest with yourself, they won't ground out into "and therefore, no matter what I do, everything is going to be OK." No matter how hard you try to justify your reasoning, the meta-reasoning tower does not terminate at "and thus, eventually you will become capable of success." They terminate at "I may be so wrong that I can never be corrected; I may fail and all value may be lost." You will find no objectively stable perch from which to launch your reasoning.

But you were *created already in motion.* You don't *need* to ground out all your beliefs and justify all your reasoning steps before you can start moving. You don't need to have plans for every contingency before you can act. You don't need to be highly confident in your analyses before you present a model. If you sit around awaiting certainty, you will be waiting a long while.

Better, I say, to cover each lack of confidence on one level with confidence on the next level, and to come to terms with the fact that if you're so irredeemable that even your best meta-reasoning cannot save you, then you've already lost.

Desperation

This is a linkpost for https://mindingourway.com/desperation/

The next three posts will discuss what I dub the three dubious virtues: desperation, recklessness, and defiance. I call them dubious, because each can easily turn into a vice if used incorrectly or excessively. As you read these posts, keep in mind the law of equal and opposite advice. Though these virtues are dubious, I have found each of them to be a crucial component of a strong and healthy intrinsic motivation system.

The first of the three dubious virtues is *desperation*. There are bad ways to be desperate: visible desperation towards *people* can put you in a bad social position, strain your relationships, or otherwise harm you. Desperation towards a *goal*, on the other hand, is vital for a guilt-free intrinsic drive.

By "desperation towards a goal" I mean the possession of a goal so important to you that you can commit yourself to it fully, without hesitation, without some part of you wondering whether it's really worth all your effort. I mean a goal that you pursue with both reckless abandon and cautious deliberation in fair portions. I mean a goal so important that it does not occur to you to spare time wondering whether you can achieve it, but only whether *this* path to achieving it is better or worse than *that* path.

In my experience, the really powerful intrinsic motivations require that you're able to struggle as if something of incredible value is on the line. That's much easier if, on a gut level, you believe that's true.

Desperate people have a power that others lack: they have the *ability to go all out,* to put all their effort towards a task without reservation. Most people I have met don't have the ability to go all out for *anything,* not even in their imagination.

Ask yourself: is there anything *you* would go all out for? Is there anything some antagonist could put in danger, such that you would pull out all your stops? Is there any threat so dire that you would hold nothing back, in your struggle to make things right?

I have met many people who cannot honestly answer "yes" to this question, not even under imaginary circumstances. If I ask them to imagine their family being kidnapped, they say they would call the police and wait anxiously. If I ask them to imagine the world threatened by an asteroid, they say they would do their best to enjoy their remaining time. These are fine and prudent answers. Yet, even if I ask them to imagine strange scenarios where they and they alone can save the Earth at great personal cost, they often say they would do it only grudgingly.

For example, imagine that aliens that want to toy with *you in particular* have put a black hole on a collision course with Earth. Imagine that the only way to redirect it is using alien tech on an alien space ship that has been left on Earth and which can be piloted only by you and you alone — and that, to destroy the black hole, you must cross the event horizon, never to return. Would you save the world then? And if so, would you do it only grudgingly?

Would you do it if the spacecraft was sequestered atop Mt. Everest? How hard would you struggle to get to the ship, if it was at the bottom of the ocean? What if it could

only be operated if you spoke fluent Mandarin, and you only had one year to learn?

Would you go all out to save the world, or would you put in a token "best effort", a token "at least I tried", and then go back to enjoying your remaining time?

And if you can't go all out even in incredible imaginary scenarios where everything depends on you, *what are you holding out for?*

A common protest here goes "I don't want to lose my friendships, my close connections, my comfort. That is too high a price to pay. If the struggle would be too brutal, then I would prefer to enjoy my remaining time instead." But if that were the case, then why couldn't someone get you to go all out by putting your friendships, connections, and comfort on the line? Would you fight with everything you have for *those*? And if not, *what are you holding out for?*

Why are you stopping yourself from putting in a full effort, if there is no situation even in principle which could compel you to pull out all the stops? Why are you holding part of yourself back, if there is nothing *even in imagination* for which you would unbar all the holds? If there is nothing anyone could put on the line such that you'd struggle with all of your being, then *what are you holding out for?*

I'm not saying you need to be willing to go all out for something *real.* It may be that the only scenarios where you'd really struggle for all you're worth are fanciful or ridiculous. I'm saying that you need to be able to go all out in principle.

There's a certain type of vulnerability that comes with committing your whole self to something. Our culture has strong social stigmas against people who *really unabashedly care* about something.

I remember a classmate in gradeschool who *really really cared* about Pokemon, to the point that others felt embarrassed just to associate with him. The stereotypical stigma against "nerds" seems rooted at least partially in a stigma against caring too much. Derision among the intellectual elites towards people who get really interested in sports seems to draw at least partially on the same stigma.

Notice the negative connotations attached to words like "cultist", "zealot", and "idealist". Notice all the people who distance themselves from whatever social movement they're in; those people who loosely identify as "effective altruists" or "rationalist" or "skeptics" or "atheists" but feel a deep compulsion to make sure you know that they think the *other* EAs/rationalists/skeptics/atheists are naive, Doing It Wrong, and blinded by their lack of nuanced views. I think that this is, in part, an attempt to defend against the curse of Caring Too Much.

Caring hard is uncool. The stereotypical intellectual is a detached moral non-realist who understands that nothing really matters, and looks upon all those "caring" folk with cynical bemusement.

Caring hard is vulnerable. If you care hard about something, then it becomes possible to lose something very important to you. Worse, everyone around you might think that you're putting your caring into the *wrong* thing, and see you as one of the naive blind idealist sheeple, and curl their lips at you.

Desperation is about *none of that mattering.* It's about having a goal so important that the social concerns drop away, except exactly insofar as they're relevant to the

achievement of your goal. It's about being willing to let yourself care more about the task at hand than about what everyone thinks about you, no matter how much they would deride you for fully committing.

A common barrier to desperation is that it can be difficult to admit that you really, really care about something, because then that means you are vulnerable to the loss of something that's very, very important to you. If your desperation is visible in a hostile social environment, desperation can destroy your ability to bargain and put you at a social disadvantage. Being social creatures, I suspect that many of us have mental architectures that prevent us from feeling desperation, because if we felt it, we'd show it, and that would undermine our social standing. (In my experience, confidence all the way up helps alleviate this effect.)

Thus, if you want to make desperation part of your intrinsic drive, you may need to practice becoming able to admit, to yourself, on a gut level, that you might lose something so terribly important that it's worth gaining a little desperation. You must first *allow yourself to become desperate.* (This is why I wrote about seeing the dark world and coming to your terms before writing about desperation.)

There is a common failure mode among those who succeed at becoming desperate, which is that they burn their resources too quickly, in their desperation. If you have to get yourself into an alien spacecraft at the bottom of the ocean, and it's going to take many months of training, social and political maneuvering, and monotonous searching, then you would be unwise to spend your first week all wound up at maximum stress levels simply because you think that that's what it means to "go all out" and "hold nothing back." If you're going to pull out all the stops and unbar all the holds, you need to understand how to carry on a slow burn as well as a fast burn. (This is why I wrote about how to avoid working yourself ragged and rest in motion before writing about desperation.)

With these tools in hand, I suggest finding a way to *become able to become desperate.* Perform whatever thought experiments and meditations you have to to be able to *imagine a situation* where you would do everything in your power to achieve some outcome, without regard for the consequences (beyond their affect on the outcome). Figure out the circumstances under which you'd pull out all the stops and unbar all the holds and put *everything you have* into the struggle.

(If there is no situation, even in theory, where you would give everything you have into your efforts, then consider that there may be a part of yourself that you're holding back for nothing, a part of yourself that you're wasting.)

I'm not saying you need to become desperate *now*. That may be unnecessary. Maybe your life is going well enough, and your goals are well enough achieved, that the best way to continue achieving them is to strengthen your friendships and your connections and enjoy your comforts. If your family *is* kidnapped, you probably *would* do best to call the police and then wait anxiously. If Earth *is* threatened by an asteroid, most people *would* do best to leave it to the experts and enjoy what time they have. So be it not upon me to force desperation upon you if you're leading a comfortable life. Make sure you don't suffer from the listless guilt, and make sure you can in principle *become* desperate, so as to ensure that you're not holding a part of yourself back for nothing, but save the actual desperation for times of need.

If, on the other hand, you are in a time of need, if you're the sort who sees every death as a tragedy, if you're otherwise fighting for something larger than yourself, then *get desperate now.*

The first step is allowing yourself to become desperate in principle. It's allowing there to be at least one imaginary scenario where you'd let yourself commit fully to a task without hesitation. Once you are able to do this, imagine the feeling that would come over you when you first committed yourself to that crucial undertaking, come whatever may. Is there a sense of desperation you would feel, a grasping need to *change the future?* Sit with it, become familiar with the sensation of desperation and any other feelings associated with the imaginary commitment.

Once you've gained some familiarity with those feelings, look with fresh eyes at what you're fighting for, at what you have to protect, at what you value, and see if any of it is worthy of a little desperation.

The art of response

This is a linkpost for https://mindingourway.com/the-art-of-response/

Imagine two different software engineers in job interviews. Both are asked for an algorithm that solves some programming puzzle, such as "identify all palindromes in a string of characters."

The first candidate, Alice, reflexively enters problem-solving mode upon hearing the problem. She pauses for a few seconds as she internalizes the problem, and then quickly thinks up a very inefficient algorithm that finds the answer by brute force. She decides to sketch this algorithm first (as a warm up) and then turn her mind to finding a more efficient path to the answer.

The second candidate, Bob, responds very differently to the same problem. He reflexively predicts that he won't be able to solve the problem. He struggles to quiet that voice in his head while he waits for a solution to present itself, but no solution is forthcoming. He struggles to focus as the seconds pass, until a part of his brain points out that he's been quiet for an uncomfortably long time, and the interviewer probably already thinks he's stupid. From then on, his thoughts are stuck on the situation, despite his attempts to wrest them back to the task at hand.

Part of what makes the difference between Alice and Bob might be skill: Alice might have more experience that lets her solve programming puzzles with less concerted effort, which helps her get to a solution before self-doubt creeps in. Self-confidence may also be a factor: perhaps Alice is simply less prone to self-doubt, and therefore less prone to this type of self-sabotage.

A third difference between Alice and Bob is their *response pattern.* Bob begins by waiting blankly for a solution to present itself; Alice begins by checking whether she can solve a simple version of the problem ("can I solve it by brute force?"). Bob is more liable to panic when no answer comes ("I have been quiet for too long"), Alice is more liable to break the problem down further if no solution presents itself ("Can I divide and conquer?").

This difference is also explained in part by experience: a more seasoned software engineer is more likely to *reflexively* notice that a problem can be solved with a simple recursion, and know which data structures to apply where. I don't think it's *only* experience, though. Imagine Alice and Bob both faced with a second problem, outside their usual comfort zone — say, a friend asks them for advice about how to handle a major life-changing event. It's easy to imagine Alice attempting to understand the situation better and asking clarifying questions that help her understand how her friend is thinking about the question. It's similarly easy to imagine Bob feeling profoundly uncomfortable, while he tries to give neutral advice and worries about the fact that he might give bad advice that ruins his friend's life.

One might call what Alice is doing "confidence," but that doesn't tell us *how it's working.* And 'confidence' also comes with connotations that may not apply to Alice — she may well decide that she isn't in a position to give good advice, she may be working from a shaky understanding and thus doubt her own conclusions, even as she turns her thoughts to understanding the obstacle before her.

One of the big differences, as I see it, is the difference in the *response pattern* between Alice and Bob. Alice justs *gets down to addressing the obstacle before her*, Bob spends mental cycles floundering. Managing response patterns is something of an art: when confronted with an obstacle, does your brain switch into problem-solving gear or do you start to flail?

Note that the art of response is *not* about immediately solving any problem placed before you. Sometimes, the best automatic response is to find some way to disengage or dodge. You aren't obligated to solve every problem placed before you. The goal of having appropriate response patterns is to *avoid flailing* and *avoid staring blankly*. The goal is to have your mind shift into the problem-solving gear.

Having effective responses prepared isn't necessarily a general skill. I'm a computer programmer at heart, and a few years ago I switched paths to math research. If I'm faced with a programming problem that I want to solve, I quickly and easily slip into effective-response-mode; I can often find solutions to problems reflexively, and when I can't, I reflexively examine the problem from many different viewpoints and start breaking it down. Yet, if you confront me with a math problem I want solved, there are still times when my reflexive response is to sit back and wait for someone else to solve it for me. (It doesn't help that I'm surrounded by brilliant mathematicians who can do so successfully.) That reflexive response — the one of blanking my mind, curious while I wait for someone else to find the answer — is not a very effective response.

Effective responses aren't about answering *quickly*, either. When paired with expertise and familiarity an effective response to an obstacle will often lead to a fast answer, but oftentimes the most effective response is to pause and think. Plenty of people have very ineffective response patterns that involve opening their mouths the moment you ask them to help you confront an obstacle. Some people reflexively start solving the wrong problem, others reflexively start making excuses for themselves, still others reflexively share personal anecdotes that paint them in a positive light. Effective response patterns are not about answering fast, they're about answering *well.*

The most competent people that I know are, almost universally, people who have very effective response patterns to obstacles in their areas of expertise. The good programmers I meet reflexively start breaking a problem down the moment they decide to solve it. The stellar mathematicians I know reflexively start prodding at problems with various techniques, or reflexively identify parts of the problem that they don't yet understand. The best businesspeople among my advisors are people who listen to me describe the choice before me, and reflexively describe the costs, constraints, and opportunities they observe. Each has acquired a highly effective response pattern to problems that fall within their area of expertise. This response pattern allows them to hit an obstacle and start taking it apart, with an Alice-like mindset, rather than flailing and doubting themselves as per Bob.

Confidence, practice, and talent all help develop these specific response patterns quite a bit. That said, you can often learn someone's *response patterns* with much less effort than it takes to learn their skills: you can start thinking in terms of incentives, opportunity costs, and markets long before you become a master economist (though reading a microeconomics textbook surely doesn't hurt). Competence isn't just about believing in your capabilities; it's also about having a pattern prepared that takes you

directly to the "break down the problem and gnaw on the parts" stage without ever dumping you into the "worry about how you've been silent for a long time and reflect on the fact that the interviewer probably thinks you're dumb" zone.

Having an *explicit* pattern, such as a checklist, can help you switch from one pattern to the other. For example, imagine Bob in the example above had a checklist which read as follows:

> *If I start dwelling on how likely I am to fail, I will do the following. (1) Say "hmm, let me think for a few minutes" aloud. (2) Verify that I understand the problem, and ask clarifying questions if I don't. (3) Check whether I could easily solve the problem by brute force. (3) Come up with a few simplifications of the problem. (4) Find a way to break off only one part of the problem or one of its simplified variants.*

then he may well be able to manually switch from a flailing response pattern to an effective one. This sort of manual switching is a good way to instill a new response pattern. The ultimate goal, though, is for efficient response patterns to become *reflexive*.

In fact, I think many people could benefit from developing efficient "fallback" response patterns, to handle new or surprising situations. Response patterns like "verify that your observations were correct" or "find more data" or "generate more than one plausible explanation for the surprise" and so on. As far as I can tell, there *is* a general skill of being able to smoothly handle surprising new situations and think on your feet, and I suspect this can be attained by developing good response patterns designed for surprising new situations.

This advice is not new, of course. Lots of self-help advice will tell you to break down the problems before you into smaller parts, and to infuse your actions with intentionality, and to reflexively do the obvious things, and so on. So I won't say much more on how to *attain* the Alice-like mindstate as opposed to the Bob-like mindstate. The important takeaway is that sometimes people respond to obstacles by breaking them down and other times they respond by flailing, and one way or another, it's useful to develop reflexive responses that put you into the former mindstate.

The way that I do it is by monitoring the ways that I respond to new obstacles placed before me. I watch myself facing various situations and observe which ones lead be to reflexively get defensive, or to reflexively blank my mind and wait for someone else to answer, or to reflexively freeze in shock and act dumbfounded. Then I practice building better response patterns for those situations, by figuring out what the checklists to run are, and I do my best to replace those patterns with reflexive inquiry, curiosity, requests for clarification, and impulses to take initiative. Polished response patterns have proven useful to me, and I attribute much of my skill at math, programming, and running nonprofits to having sane responses to new obstacles.

Regardless of where you get your response patterns from, I suspect that honing them will do you well.