Best of LessWrong: January 2014

1. Dark Arts of Rationality
2. Tell Culture
3. A big Singularity-themed Hollywood movie out in April offers many opportunities to talk about AI risk
4. On saving the world
5. The mechanics of my recent productivity
6. Dangers of steelmanning / principle of charity
7. Rationalists Are Less Credulous But Better At Taking Ideas Seriously
8. Handshakes, Hi, and What's New: What's Going On With Small Talk?
9. Deregulating Distraction, Moving Towards the Goal, and Level Hopping
10. Dr. Jubjub predicts a crisis
11. LessWrong Help Desk - free paper downloads and more (2014)
12. Try more things.
13. Habitual Productivity
14. Results from MIRI's December workshop
15. Book Review: How Learning Works
16. Fascists and Rakes
17. Even Odds
18. Humans can drive cars
19. [link] Why Self-Control Seems (but may not be) Limited
20. Soft Paternalism in Parenting
21. Decision Auctions aka "How to fairly assign chores, or decide who gets the last cookie"
22. European Community Weekend in Berlin
23. Things I Wish They'd Taught Me When I Was Younger: Why Money Is Awesome
24. I Will Pay $500 To Anyone Who Can Convince Me To Cancel My Cryonics Subscription
25. Literature-review on cognitive effects of modafinil (my bachelor thesis)

Best of LessWrong: January 2014

1. Dark Arts of Rationality
2. Tell Culture
3. A big Singularity-themed Hollywood movie out in April offers many opportunities to talk about AI risk
4. On saving the world
5. The mechanics of my recent productivity
6. Dangers of steelmanning / principle of charity
7. Rationalists Are Less Credulous But Better At Taking Ideas Seriously
8. Handshakes, Hi, and What's New: What's Going On With Small Talk?
9. Deregulating Distraction, Moving Towards the Goal, and Level Hopping
10. Dr. Jubjub predicts a crisis
11. LessWrong Help Desk - free paper downloads and more (2014)
12. Try more things.
13. Habitual Productivity
14. Results from MIRI's December workshop
15. Book Review: How Learning Works
16. Fascists and Rakes
17. Even Odds
18. Humans can drive cars
19. [link] Why Self-Control Seems (but may not be) Limited
20. Soft Paternalism in Parenting
21. Decision Auctions aka "How to fairly assign chores, or decide who gets the last cookie"
22. European Community Weekend in Berlin
23. Things I Wish They'd Taught Me When I Was Younger: Why Money Is Awesome
24. I Will Pay $500 To Anyone Who Can Convince Me To Cancel My Cryonics Subscription
25. Literature-review on cognitive effects of modafinil (my bachelor thesis)

Dark Arts of Rationality

Today, we're going to talk about Dark rationalist techniques: productivity tools which seem incoherent, mad, and downright irrational. These techniques include:

1. Willful Inconsistency
2. Intentional Compartmentalization
3. Modifying Terminal Goals

I expect many of you are already up in arms. It seems obvious that consistency is a virtue, that compartmentalization is a flaw, and that one should *never* modify their terminal goals.

I claim that these 'obvious' objections are incorrect, and that all three of these techniques can be instrumentally rational.

In this article, I'll promote the strategic cultivation of false beliefs and condone mindhacking on the values you hold most dear. Truly, these are Dark Arts. I aim to convince you that sometimes, the benefits are worth the price.

Changing your Terminal Goals

In many games there is no "absolutely optimal" strategy. Consider the Prisoner's Dilemma. The optimal strategy depends entirely upon the strategies of the other players. *Entirely.*

Intuitively, you may believe that there are some fixed "rational" strategies. Perhaps you think that even though complex behavior is dependent upon other players, there are still *some* constants, like "Never cooperate with DefectBot". DefectBot always defects against you, so you should never cooperate with it. Cooperating with DefectBot would be insane. Right?

Wrong. If you find yourself on a playing field where everyone else is a TrollBot (players who cooperate with you if and only if you cooperate with DefectBot) then you should cooperate with DefectBots and defect against TrollBots.

Consider that. There are playing fields where you should *cooperate with DefectBot*, even though that looks completely insane from a naïve viewpoint. Optimality is not a feature of the strategy, it is a relationship between the strategy and the playing field.

Take this lesson to heart: in certain games, there are strange playing fields where the optimal move looks *completely irrational*.

I'm here to convince you that *life* is one of those games, and that you occupy a strange playing field *right now*.

Here's a toy example of a strange playing field, which illustrates the fact that even your terminal goals are not sacred:

Imagine that you are completely self-consistent and have a utility function. For the sake of the thought experiment, pretend that your terminal goals are distinct,

exclusive, orthogonal, and clearly labeled. You value your goals being achieved, but you have no preferences about *how* they are achieved or what happens afterwards (unless the goal explicitly mentions the past/future, in which case achieving the goal puts limits on the past/future). You possess at least two terminal goals, one of which we will call A.

Omega descends from on high and makes you an offer. Omega will cause your terminal goal A to become achieved over a certain span of time, without any expenditure of resources. As a price of taking the offer, you must switch out terminal goal A for terminal goal B. Omega guarantees that B is orthogonal to A and all your other terminal goals. Omega further guarantees that you will achieve B using less time and resources than you would have spent on A. Any other concerns you have are addressed via similar guarantees.

Clearly, you should take the offer. One of your terminal goals will be achieved, and while you'll be pursuing a new terminal goal that you (before the offer) don't care about, you'll come out ahead in terms of time and resources which can be spent achieving your other goals.

So the optimal move, in this scenario, is to change your terminal goals.

There are times when the optimal move of a rational agent is to hack its own terminal goals.

You may find this counter-intuitive. It helps to remember that "optimality" depends as much upon the playing field as upon the strategy.

Next, I claim that such scenarios not restricted to toy games where Omega messes with your head. Humans encounter similar situations on a day-to-day basis.

Humans often find themselves in a position where they should modify their terminal goals, and the reason is simple: our thoughts do not have direct control over our motivation.

Unfortunately for us, our "motivation circuits" can distinguish between terminal and instrumental goals. It is often easier to put in effort, experience inspiration, and work tirelessly when pursuing a terminal goal as opposed to an instrumental goal. It would be nice if this were not the case, but it's a *fact of our hardware*: we're going to do X more if we want to do X for its own sake as opposed to when we force X upon ourselves.

Consider, for example, a young woman who wants to be a rockstar. She wants the fame, the money, and the lifestyle: these are her "terminal goals". She lives in some strange world where rockstardom is wholly dependent upon merit (rather than social luck and network effects), and decides that in order to become a rockstar she has to produce really good music.

But here's the problem: She's a human. Her conscious decisions don't directly affect her motivation.

In her case, it turns out that she can make better music when "Make Good Music" is a terminal goal as opposed to an instrumental goal.

When "Make Good Music" is an instrumental goal, she schedules practice time on a sitar and grinds out the hours. But she doesn't really *like* it, so she cuts corners whenever akrasia comes knocking. She lacks inspiration and spends her spare hours dreaming of stardom. Her songs are shallow and trite.

When "Make Good Music" is a terminal goal, music pours forth, and she spends every spare hour playing her sitar: not because she knows that she "should" practice, but because you couldn't pry her sitar from her cold dead fingers. She's not "practicing", she's pouring out her soul, and no power in the 'verse can stop her. Her songs are emotional, deep, and moving.

It's obvious that she should adopt a new terminal goal.

Ideally, we would be just as motivated to carry out instrumental goals as we are to carry out terminal goals. In reality, this is not the case. As a human, your motivation system *does* discriminate between the goals that you feel obligated to achieve and the goals that you pursue as ends unto themselves.

As such, it is sometimes in your best interest to modify your terminal goals.

Mind the terminology, here. When I speak of "terminal goals" I mean actions that feel like ends unto themselves. I am speaking of the stuff you wish you were doing when you're doing boring stuff, the things you do in your free time just because they are *fun*, the actions you don't need to justify.

This seems like the obvious meaning of "terminal goals" to me, but some of you may think of "terminal goals" more akin to self-endorsed morally sound end-values in some consistent utility function. I'm not talking about those. I'm not even convinced I have any.

Both types of "terminal goal" are susceptible to strange playing fields in which the optimal move is to change your goals, but it is only the former type of goal — the actions that are simply *fun*, that need no justification — which I'm suggesting you tweak for instrumental reasons.

I've largely refrained from goal-hacking, personally. I bring it up for a few reasons:

1. It's the easiest Dark Side technique to justify. It helps break people out of the mindset where they think optimal actions are the ones that look rational in a vacuum. Remember, optimality is a feature of the playing field. Sometimes cooperating with DefectBot is the best strategy!
2. Goal hacking segues nicely into the other Dark Side techniques which I use frequently, as you will see shortly.
3. I have met many people who would benefit from a solid bout of goal-hacking.

I've crossed paths with many a confused person who (without any explicit thought on their part) had really silly terminal goals. We've all met people who are acting as if "Acquire Money" is a terminal goal, never noticing that money is almost entirely instrumental in nature. When you ask them "but what would you do if money was no issue and you had a lot of time", all you get is a blank stare.

Even the LessWrong Wiki entry on terminal values describes a college student for which university is instrumental, and getting a job is terminal. This seems like a clear-

cut case of a Lost Purpose: a job seems clearly instrumental. And yet, we've all met people who act as if "Have a Job" is a terminal value, and who then seem aimless and undirected after finding employment.

These people could use some goal hacking. You can argue that Acquire Money and Have a Job aren't "really" terminal goals, to which I counter that many people don't know their ass from their elbow when it comes to their own goals. Goal hacking is an important part of becoming a rationalist and/or improving mental health.

Goal-hacking in the name of consistency isn't really a Dark Side power. This power is only Dark when you use it like the musician in our example, when you adopt terminal goals for instrumental reasons. This form of goal hacking is less common, but can be very effective.

I recently had a personal conversation with Alexei, who is earning to give. He noted that he was not entirely satisfied with his day-to-day work, and mused that perhaps goal-hacking (making "Do Well at Work" an end unto itself) could make him more effective, generally happier, and more productive in the long run.

Goal-hacking can be a powerful technique, when correctly applied. Remember, you're not in direct control of your motivation circuits. Sometimes, strange though it seems, the optimal action involves fooling *yourself*.

You don't get good at programming by sitting down and forcing yourself to practice for three hours a day. I mean, I suppose you *could* get good at programming that way. But it's much easier to get good at programming by *loving programming*, by being the type of person who spends every spare hour tinkering on a project. Because then it doesn't feel like practice, it feels like fun.

This is the power that you can harness, if you're willing to tamper with your terminal goals for instrumental reasons. As rationalists, we would prefer to dedicate to instrumental goals the same vigor that is reserved for terminal goals. Unfortunately, we find ourselves on a strange playing field where goals that feel justified in their own right win the lion's share of our attention.

Given this strange playing field, goal-hacking can be optimal.

You don't have to completely mangle your goal system. Our aspiring musician from earlier doesn't need to destroy her "Become a Rockstar" goal in order to adopt the "Make Good Music" goal. If you can successfully convince yourself to believe that something instrumental is a means unto itself (e.g. terminal), *while still believing that it is instrumental*, then more power to you.

This is, of course, an instance of Intentional Compartmentalization.

Intentional Compartmentalization

As soon as you endorse modifying your own terminal goals, Intentional Compartmentalization starts looking like a pretty good idea. If Omega offers to achieve `A` at the price of dropping `A` and adopting `B`, the ideal move is to take the offer after finding a way to not *actually* care about `B`.

A consistent agent cannot do this, but I have good news for you: You're a human. You're not consistent. In fact, you're *great* at being inconsistent!

You might expect it to be difficult to add a new terminal goal while still believing that it's instrumental. You may also run into strange situations where holding an instrumental goal as terminal *directly contradicts* other terminal goals.

For example, our aspiring musician might find that she makes even *better* music if "Become a Rockstar" is *not* among her terminal goals.

This means she's in trouble: She either has to drop "Become a Rockstar" and have a better chance at *actually becoming a rockstar*, or she has to settle for a decreased chance that she'll become a rockstar.

Or, rather, she would have to settle for one of these choices — if she wasn't human.

I have good news! Humans are *really really* good at being inconsistent, and you can leverage this to your advantage. Compartmentalize! Maintain goals that are "terminal" in one compartment, but which you know are "instrumental" in another, then simply never let those compartments touch!

This may sound completely crazy and irrational, but remember: you aren't actually in control of your motivation system. You find yourself on a strange playing field, and the optimal move may in fact require mental contortions that make epistemic rationalists shudder.

Hopefully you never run into this particular problem (holding contradictory goals in "terminal" positions), but this illustrates that there are scenarios where compartmentalization works in your favor. Of course we'd *prefer* to have direct control of our motivation systems, but *given that we don't*, compartmentalization is a huge asset.

Take a moment and let this sink in before moving on.

Once you realize that compartmentalization is OK, you are ready to practice my second Dark Side technique: Intentional Compartmentalization. It has many uses outside the realm of goal-hacking.

See, motivation is a fickle beast. And, as you'll remember, your conscious choices are not directly attached to your motivation levels. You can't just *decide* to be more motivated.

At least, not directly.

I've found that certain beliefs — beliefs which I *know are wrong* — can make me more productive. (On a related note, remember that religious organizations are generally more coordinated than rationalist groups.)

It turns out that, under these false beliefs, I can tap into motivational reserves that are otherwise unavailable. The only problem is, I know that these beliefs are downright false.

I'm just kidding, that's not actually a problem. Compartmentalization to the rescue!

Here's a couple example beliefs that I keep locked away in my mental compartments, bound up in chains. Every so often, when I need to be extra productive, I don my protective gear and enter these compartments. I never fully believe these things — not globally, at least — but I'm capable of attaining "local belief", of acting as if I hold these beliefs. This, it turns out, is enough.

Nothing is Beyond My Grasp

We'll start off with a tame belief, something that is soundly rooted in evidence outside of its little compartment.

I have a global belief, outside all my compartments, that nothing is beyond my grasp.

Others may understand things easier I do or faster than I do. People smarter than myself grok concepts with less effort than I. It may take me *years* to wrap my head around things that other people find trivial. However, there is no idea that a human has ever had that I cannot, *in principle*, grok.

I believe this with moderately high probability, just based on my own general intelligence and the fact that brains are so tightly clustered in mind-space. It may take me a hundred times the effort to understand something, but I can still understand it eventually. Even things that are beyond the grasp of a meager human mind, I will one day be able to grasp after I upgrade my brain. Even if there are limits imposed by reality, I could *in principle* overcome them if I had enough computing power. Given any finite idea, I could in theory become powerful enough to understand it.

This belief, itself, is not compartmentalized. What is compartmentalized is the *certainty*.

Inside the compartment, I believe that Nothing is Beyond My Grasp with 100% confidence. Note that this is ridiculous: there's no such thing as 100% confidence. At least, not in my global beliefs. But inside the compartments, while we're in la-la land, it helps to treat Nothing is Beyond My Grasp as raw, immutable *fact*.

You might think that it's sufficient to believe Nothing is Beyond My Grasp with very high probability. If that's the case, you haven't been listening: I *don't* actually believe Nothing is Beyond My Grasp with an extraordinarily high probability. I believe it with moderate probability, and then I *have a compartment* in which it's a certainty.

It would be *nice* if I never needed to use the compartment, if I could face down technical problems and incomprehensible lingo and being really out of my depth with a relatively high confidence that I'm going to be able to make sense of it all. However, I'm not in direct control of my motivation. And it turns out that, through some quirk in my psychology, it's easier to face down the oppressive feeling of being in *way over my head* if I have this rock-solid "belief" that Nothing is Beyond My Grasp.

This is what the compartments are good for: I don't actually believe the things inside them, but I can still *act as if I do*. That ability allows me to face down challenges that would be difficult to face down otherwise.

This compartment was largely constructed with the help of The Phantom Tollbooth: it taught me that there are certain impossible tasks you can do if you think they're possible. It's not always enough to know that if I believe I can do a thing, then I have a

higher probability of being able to do it. I get an extra boost from believing I can do *anything*.

You might be surprised about how much you can do when you have a mental compartment in which you are *unstoppable*.

My Willpower Does Not Deplete

Here's another: My Willpower Does Not Deplete.

Ok, so my willpower actually does deplete. I've been writing about how it does, and discussing methods that I use to avoid depletion. *Right now*, I'm writing about how I've acknowledged the fact that my willpower *does deplete*.

But I have this compartment where it doesn't.

Ego depletion is a funny thing. If you don't believe in ego depletion, you suffer less ego depletion. This does not eliminate ego depletion.

Knowing this, I have a compartment in which My Willpower Does Not Deplete. I go there often, when I'm studying. It's easy, I think, for one to begin to feel tired, and say "oh, this must be ego depletion, I can't work anymore." Whenever my brain tries to go there, I wheel this bad boy out of his cage. "Nope", I respond, "My Willpower Does Not Deplete".

Surprisingly, this often works. I won't force myself to keep working, but I'm pretty good at preventing mental escape attempts via "phantom akrasia". I don't allow myself to invoke ego depletion or akrasia to stop being productive, because My Willpower Does Not Deplete. I have to *actually be tired out*, in a way that doesn't trigger the My Willpower Does Not Deplete safeguards. This doesn't let me keep going forever, but it prevents a lot of false alarms.

In my experience, the strong version (My Willpower Does Not Deplete) is much more effective than the weak version (My Willpower is Not Depleted Yet), even though it's more wrong. This probably says something about my personality. Your mileage may vary. Keep in mind, though, that the effectiveness of your mental compartments may depend more on the motivational content than on degree of falsehood.

Anything is a Placebo

Placebos work even when you know they are placebos.

This is the sort of madness I'm talking about, when I say things like "you're on a strange playing field".

Knowing this, you can easily activate the placebo effect manually. Feeling sick? Here's a freebie: drink more water. It will make you feel better.

No? It's just a placebo, you say? Doesn't matter. Tell yourself that water makes it better. Put that in a nice little compartment, save it for later. It doesn't matter that you know what you're doing: your brain is easily fooled.

Want to be more productive, be healthier, and exercise more effectively? Try using Anything is a Placebo! Pick something trivial and non-harmful and tell yourself that it helps you perform better. Put the belief in a compartment in which you *act as if* you believe the thing. Cognitive dissonance doesn't matter! Your brain is *great* at ignoring cognitive dissonance. You can "know" you're wrong in the global case, while "believing" you're right locally.

For bonus points, try combining objectives. Are you constantly underhydrated? Try believing that drinking more water makes you more alert!

Brains are weird.

Truly, these are the Dark Arts of instrumental rationality. Epistemic rationalists recoil in horror as I advocate *intentionally cultivating false beliefs.* It goes without saying that you should use this technique with care. Remember to always audit your compartmentalized beliefs through the lens of your actual beliefs, and be very careful not to let incorrect beliefs leak out of their compartments.

If you think you can achieve similar benefits without "fooling yourself", then by all means, do so. I haven't been able to find effective alternatives. Brains have been honing compartmentalization techniques for *eons*, so I figure I might as well re-use the hardware.

It's important to reiterate that these techniques are necessary because *you're not actually in control of your own motivation*. Sometimes, incorrect beliefs make you more motivated. Intentionally cultivating incorrect beliefs is surely a path to the Dark Side: compartmentalization only mitigates the damage. If you make sure you segregate the bad beliefs and acknowledge them for what they are then you can get much of the benefit without paying the cost, but there is still a cost, and the currency is cognitive dissonance.

At this point, you should be mildly uncomfortable. After all, I'm advocating something which is completely epistemically irrational. We're not done yet, though.

I have one more Dark Side technique, and it's worse.

Willful Inconsistency

I use Intentional Compartmentalization to "locally believe" things that I don't "globally believe", in cases where the local belief makes me more productive. In this case, the beliefs in the compartments are things that I tell myself. They're like mantras that I repeat in my head, at the System 2 level. System 1 is fragmented and compartmentalized, and happily obliges.

Willful Inconsistency is the grown-up, scary version of Intentional Compartmentalization. It involves convincing System 1 wholly and entirely of something that System 2 does not actually believe. There's no compartmentalization and no fragmentation. There's nowhere to shove the incorrect belief when you're done with it. It's taken over the intuition, and it's always on. Willful Inconsistency is about having gut-level intuitive beliefs that you explicitly disavow.

Your intuitions run the show whenever you're not paying attention, so if you're willfully inconsistent then you're going to actually *act as if* these incorrect beliefs are true in your day-to-day life, unless your forcibly override your default actions. Ego depletion and distraction make you vulnerable *to yourself*.

Use this technique with caution.

This may seem insane even to those of you who took the previous suggestions in stride. That you must sometimes alter your terminal goals is a feature of the playing field, not the agent. The fact that you are not in direct control of your motivation system readily implies that tricking yourself is useful, and compartmentalization is an obvious way to mitigate the damage.

But why would anyone ever try to convince themselves, deep down at the core, of something that they don't actually believe?

The answer is simple: specialization.

To illustrate, let me explain how I use willful inconsistency.

I have invoked Willful Inconsistency on only two occasions, and they were similar in nature. Only one instance of Willful Inconsistency is currently active, and it works like this:

I have completely and totally convinced my intuitions that unfriendly AI is a problem. A big problem. System 1 operates under the assumption that UFAI will come to pass in the next twenty years with very high probability.

You can imagine how this is somewhat motivating.

On the conscious level, within System 2, I'm much less certain. I solidly believe that UFAI is a big problem, and that it's the problem that I should be focusing my efforts on. However, my error bars are *far* wider, my timespan is quite broad. I acknowledge a decent probability of soft takeoff. I assign moderate probabilities to a number of other existential threats. I think there are a large number of unknown unknowns, and there's a non-zero chance that the status quo continues until I die (and that I can't later be brought back). All this I know.

But, *right now*, as I type this, my intuition is screaming at me that the above is all wrong, that my error bars are narrow, and that I don't *actually* expect the status quo to continue for even thirty years.

This is just how I like things.

See, I *am* convinced that building a friendly AI is the most important problem for me to be working on, *even though* there is a very real chance that MIRI's research won't turn out to be crucial. Perhaps other existential risks will get to us first. Perhaps we'll get brain uploads and Robin Hanson's emulation economy. Perhaps it's going to take far longer than expected to crack general intelligence. However, after much reflection I have concluded that despite the uncertainty, this is where I should focus my efforts.

The problem is, it's hard to translate that decision down to System 1.

Consider a toy scenario, where there are ten problems in the world. Imagine that, in the face of uncertainty and diminishing returns from research effort, I have concluded

that the world should allocate 30% of resources to problem A, 25% to problem B, 10% to problem C, and 5% to each of the remaining problems.

Because specialization leads to massive benefits, it's much more effective to dedicate 30% of researchers to working on problem A rather than having all researchers dedicate 30% of their time to problem A. So presume that, in light of these conclusions, I decide to dedicate myself to problem A.

Here we have a problem: I'm supposed to specialize in problem A, but at the intuitive level problem A isn't *that* big a deal. It's only 30% of the problem space, after all, and it's not really that much worse than problem B.

This would be no issue if I were in control of my own motivation system: I could put the blinders on and focus on problem A, crank the motivation knob to maximum, and trust everyone else to focus on the other problems and do their part.

But I'm not in control of my motivation system. If my intuitions know that there are a number of other similarly worthy problems that I'm ignoring, if they are distracted by other issues of similar scope, then I'm tempted to work on everything at once. This is bad, because output is maximized if we all specialize.

Things get especially bad when problem A is highly uncertain and unlikely to affect people for decades if not centuries. It's very hard to convince the monkey brain to care about far-future vagaries, *even if* I've rationally concluded that those are where I should dedicate my resources.

I find myself on a strange playing field, where the optimal move is to lie to System 1.

Allow me to make that more concrete:

I'm *much* more motivated to do FAI research when I'm intuitively convinced that we have a hard 15 year timer until UFAI.

Explicitly, I believe UFAI is one possibility among many and that the timeframe should be measured in decades rather than years. I've concluded that it is my most pressing concern, but I don't *actually* believe we have a hard 15 year countdown.

That said, it's hard to understate how useful it is to have a gut-level feeling that there's a short, hard timeline. This "knowledge" pushes the monkey brain to go all out, no holds barred. In other words, this is the method by which I convince myself to *actually* specialize.

This is how I convince myself to deploy every available resource, to attack the problem as if the stakes were incredibly high. Because the stakes *are* incredibly high, and I *do* need to deploy every available resource, even if we don't have a hard 15 year timer.

In other words, Willful Inconsistency is the technique I use to force my intuition to *feel as if* the stakes are as high as I've calculated them to be, given that my monkey brain is bad at responding to uncertain vague future problems. Willful Inconsistency is my counter to Scope Insensitivity: my intuition has difficulty believing the results when I do the multiplication, so I lie to it until it acts with appropriate vigor.

This is the final secret weapon in my motivational arsenal.

I don't personally recommend that you try this technique. It can have harsh side effects, including feelings of guilt, intense stress, and massive amounts of cognitive dissonance. I'm able to do this in large part because I'm in a very good headspace. I went into this with full knowledge of what I was doing, and I am confident that I can back out (and actually correct my intuitions) if the need arises.

That said, I've found that cultivating a gut-level feeling that what you're doing *must* be done, and must be done *quickly*, is an extraordinarily good motivator. It's such a strong motivator that I seldom explicitly acknowledge it. I don't need to mentally invoke "we have to study or the world ends". Rather, this knowledge lingers in the background. It's not a mantra, it's not something that I repeat and wear thin. Instead, it's this gut-level drive that sits underneath it all, that makes me strive to go faster unless I explicitly try to slow down.

This monkey-brain tunnel vision, combined with a long habit of productivity, is what keeps me Moving Towards the Goal.

Those are my Dark Side techniques: Willful Inconsistency, Intentional Compartmentalization, and Terminal Goal Modification.

I expect that these techniques will be rather controversial. If I may be so bold, I recommend that discussion focus on goal-hacking and intentional compartmentalization. I acknowledge that willful inconsistency is unhealthy and I don't generally recommend that others try it. By contrast, both goal-hacking and intentional compartmentalization are quite sane and, indeed, instrumentally rational.

These are certainly not techniques that I would recommend CFAR teach to newcomers, and I remind you that "it is dangerous to be half a rationalist". You can royally screw you over if you're still figuring out your beliefs as you attempt to compartmentalize false beliefs. I recommend only using them when you're sure of what your goals are and confident about the borders between your actual beliefs and your intentionally false "beliefs".

It may be surprising that changing terminal goals can be an optimal strategy, and that humans should consider adopting incorrect beliefs strategically. At the least, I encourage you to remember that there are no absolutely rational actions.

Modifying your own goals and cultivating false beliefs are useful because we live in strange, hampered control systems. Your brain was optimized with no concern for truth, and optimal performance may require self deception. I remind the uncomfortable that instrumental rationality is not about being the most consistent or the most correct, it's about *winning.* There are games where the optimal move requires adopting false beliefs, and if you find yourself playing one of those games, then you should adopt false beliefs. Instrumental rationality and epistemic rationality can be pitted against each other.

We are fortunate, as humans, to be skilled at compartmentalization: this helps us work around our mental handicaps without sacrificing epistemic rationality. Of course, we'd rather not have the mental handicaps in the first place: but you have to work with what you're given.

We *are* weird agents without full control of our own minds. We lack direct control over important aspects of ourselves. For that reason, it's often necessary to take actions that may seem contradictory, crazy, or downright irrational.

Just remember this, before you condemn these techniques: optimality is as much an aspect of the playing field as of the strategy, and humans occupy a strange playing field indeed.

Tell Culture

Followup to: Ask and Guess

Ask culture: "I'll be in town this weekend for a business trip. Is it cool if I crash at your place?" Response: “Yes“ or “no”.

Guess culture: "Hey, great news! I'll be in town this weekend for a business trip!" Response: Infer that they might be telling you this because they want something from you, conclude that they might want a place to stay, and offer your hospitality only if you want to. Otherwise, pretend you didn’t infer that.

The two basic rules of Ask Culture: 1) Ask when you want something. 2) Interpret things as requests and feel free to say "no".

The two basic rules of Guess Culture: 1) Ask for things if, and *only* if, you're confident the person will say "yes". 2) Interpret requests as expectations of "yes", and, when possible, avoid saying "no".

Both approaches come with costs and benefits. In the end, I feel pretty strongly that Ask is superior.

But these are not the only two possibilities!

"I'll be in town this weekend for a business trip. I would like to stay at your place, since it would save me the cost of a hotel, plus I would enjoy seeing you and expect we’d have some fun. I'm looking for other options, though, and would rather stay elsewhere than inconvenience you." Response: “I think I need some space this weekend. But I’d love to get a beer or something while you’re in town!” or “You should totally stay with me. I’m looking forward to it.”

There is a third alternative, and I think it's probably what rationalist communities ought to strive for. I call it "Tell Culture".

The two basic rules of Tell Culture: 1) Tell the other person what's going on in your own mind whenever you suspect you'd both benefit from them knowing. (Do NOT assume others will accurately model your mind without your help, or that it will even occur to them to ask you questions to eliminate their ignorance.) 2) Interpret things people tell you as attempts to create common knowledge for shared benefit, rather than as requests or as presumptions of compliance.

Suppose you’re in a conversation that you’re finding aversive, and you can’t figure out why. Your goal is to procure a rain check.

- Guess: *You see this annoyed body language? Huh? Look at it! If you don’t stop talking soon I swear I’ll start tapping my foot.* (Or, possibly, tell a little lie to excuse yourself. “Oh, look at the time…”)
- Ask: “Can we talk about this another time?”
- Tell: "I'm beginning to find this conversation aversive, and I'm not sure why. I propose we hold off until I've figured that out."

Here are more examples from my own life:

- "I didn't sleep well last night and am feeling frazzled and irritable today. I apologize if I snap at you during this meeting. It isn’t personal."
- "I just realized this interaction will be far more productive if my brain has food. I think we should head toward the kitchen."
- "It would be awfully convenient networking for me to stick around for a bit after our meeting to talk with you and [the next person you're meeting with]. But on a scale of one to ten, it's only about 3 useful to me. If you'd rate the loss of utility for you as two or higher, then I have a strong preference for not sticking around."

The burden of honesty is even greater in Tell culture than in Ask culture. To a Guess culture person, I imagine much of the above sounds passive aggressive or manipulative, much worse than the rude bluntness of mere Ask. It’s because Guess people aren’t expecting relentless truth-telling, which is exactly what’s necessary here.

If you’re occasionally dishonest and tell people you want things you don't actually care about--like their comfort or convenience--they’ll learn not to trust you, and the inherent freedom of the system will be lost. They’ll learn that you only pretend to care about them to take advantage of their reciprocity instincts, when in fact you’ll count them as having defected if they respond by stating a preference for protecting their own interests.

Tell culture is cooperation with open source codes.

This kind of trust does not develop overnight. Here is the most useful Tell tactic I know of for developing that trust with a native Ask or Guess. It’s saved me sooooo much time and trouble, and I wish I’d thought of it earlier.

"I'm not asking because I expect you to say ‘yes’. I'm asking because I'm having trouble imagining the inside of your head, and I want to understand better. You are completely free to say ‘no’, or to tell me what you’re thinking right now, and I promise it will be fine." It is amazing how often people quickly stop looking shifty and say 'no' after this, or better yet begin to discuss further details.

A big Singularity-themed Hollywood movie out in April offers many opportunities to talk about AI risk

There's a big Hollywood movie coming out with an apocalyptic Singularity-like story, called Transcendence. (IMDB, Wiki, official site) With an A-list cast and big budget, I contend this movie is the front-runner to be 2014's most significant influence on discussions of superintelligence outside specialist circles. Anyone hoping to influence those discussions should start preparing some talking points.

I don't see anybody here agree with me on this. The movie has been briefly discussed on LW when it was first announced in March 2013, but since then, only the trailer (out since December) has been mentioned. MIRI hasn't published a word about it. This amazes me. We have three months till millions of people who never considered superintelligence are going to start thinking about it - is nobody bothering to craft a response to the movie yet? Shouldn't there be something that lazy journalists, given the job to write about this movie, can find?

Because if there isn't, they'll dismiss the danger of AI like Erik Sofge already did in an early piece about the movie for Popular Science, and nudge their readers to do so too. And that'd be a shame, wouldn't it?

On saving the world

This is the final post in my productivity sequence.

The first post described what I achieved. The next three posts describe how. This post describes why, explaining the sources of my passion and the circumstances that convinced a young Nate to try and save the world. Within, you will find no suggestions, no techniques to emulate, no new ideas to ponder. This is a rationalist coming-of-age story. With luck, you may find it inspiring. Regardless, I hope you can learn from my mistakes.

Never fear, I'll be back to business soon — there's lots of studying to do. But before then, there's a story to tell, a memorial to what I left behind.

I was raised Catholic. On my eighth birthday, having received my first communion about a year prior, I casually asked my priest how to reaffirm my faith and do something for the Lord. The memory is fuzzy, but I think I donated a chunk of allowance money and made a public confession at the following mass.

A bunch of the grownups made a big deal out of it, as grownups are like to do. "Faith of a child", and all that. This confused me, especially when I realized that what I had done was rare. I wasn't trying to get pats on the head, I was appealing to the *Lord of the Heavens and the Earth*. Were we all on the same page, here? This was the *creator.* He was infinitely virtuous, and he had told us what to do.

And yet, everyone was content to recite hymns once a week and donate for the reconstruction of the church. What about the rest of the world, the sick, the dying? Where were the proselytizers, the missionary opportunities? Why was everyone just sitting around?

On that day, I became acquainted with civilizational inadequacy. I realized you could hand a room full of people the *literal word of God*, and they'd still struggle to pay attention for an hour every weekend.

This didn't shake my faith, mind you. It didn't even occur to me that the grownups might not actually believe their tales. No, what I learned that day was that there are a lot of people who hold beliefs they aren't willing to act upon.

Eventually, my faith faded. The distrust remained.

Gaining Confidence

I grew up in a small village, population ~1200. My early education took place in a one-room schoolhouse. The local towns eventually rolled all their school districts into one, but even then, my graduating class barely broke 50 people. It wasn't difficult to excel.

Ages twelve and thirteen were rough — that was right after they merged school districts, and those were the years I was first put a few grades ahead in math classes. I was awkward and underconfident. I felt estranged and lonely, and it was easy to get shoehorned into the "smart kid" stereotype by all the new students.

Eventually, though, I decided that the stereotype was bogus. Anyone intelligent should be able to escape such pigeonholing. In fact, I concluded that anyone with real smarts should be able to find their way out of *any* mess. I observed the confidence possessed by my peers, even those who seemed to have no reason for confidence. I noticed the ease with which they engaged in social interactions. I decided I could emulate these.

I faked confidence, and it soon became real. I found that my social limitations had been largely psychological, and that the majority of my classmates were more than willing to be friends. I learned how to get good grades without alienating my peers. It helped that I tended to buck authority (I was no "teacher's pet") and that I enjoyed teaching others. I had a knack for pinpointing misunderstandings and was often able to teach better than the teachers could — as a peer, I could communicate on a different level.

I started doing very well for myself. I got excellent grades with minimal effort. I overcame my social anxieties. I had a few close friends and was on good terms with most everyone else. I participated in a number of extra circulars where I held high status. As you may imagine, I grew quite arrogant.

In retrospect, my accomplishments were hardly impressive. At the time, though, it felt like everyone else wasn't even *trying*. It became apparent that if I wanted something done right, I'd have to do it myself.

Shattered Illusions

Up until the age of fourteen I had this growing intuition that you can't trust others to actually get things done. This belief didn't become explicit until the end of ninth grade, when I learned how the government of the United States of America actually works.

Allow me to provide a few pieces of context.

For one thing, I was learning to program computers at the time. I had been programming for maybe a year and a half, and I was starting to form concepts of elegance and minimalism. I had a belief that the best design is a small design, a design forced by nature at every step along the way, a design that requires no arbitrary choices.

For another thing, my religion had died not with a bang, but with a whimper. I'd compartmentalized it, and it had slowly withered away. I didn't Believe any more, but I didn't mind that others did. It was a happy fantasy, a social tool. Just as children are allowed to believe in Santa Claus, grownups were allowed to believe in Gods.

The government, though, was a different matter all together. I assumed that a lot of very smart people had put a lot of effort into its design — that's what the "Founding Fathers" meme implied, anyway. But maybe it wasn't even that. Maybe I just possessed an unspoken, unchallenged belief that the grownups knew what they were doing, at least at the very highest levels. This was the very fabric of society itself: surely it was meticulously calibrated to maximize human virtue, to protect us from circumstance and evil.

When I was finally told how the US government worked, I couldn't believe my ears. It was a *mess*. An arbitrary, clunky monstrosity full of loopholes a child could abuse. I

could think of a dozen improvements off the top of my head.

To give you an idea of how my teenaged mind worked, it was immediately clear to me that any first-order "improvements" suggested by naïve ninth-graders would have unintended negative consequences. Therefore, improvement number one involved redesigning the system to make it easy to test many different improvements in parallel, adding machinery to adopt the improvements that were actually shown to work.

Yet even these simple ideas were absent in the actual system. Corruption and inefficiency ran rampant. Worse, my peers didn't seem particularly perturbed: they took the system as a given, and merely memorized the machinery for long enough to pass a test. Even the grownups were apathetic: they dickered over who should have power *within* the system, never suggesting we should alter the system itself.

My childhood illusions fell to pieces. I realized that nothing was meticulously managed, that the smartest people weren't in control, making sure that everything was optimal. All the world problems, the sicknesses and the injustices and the death: these weren't necessary evils, they were a product of neglect. The most important system of all was poorly coordinated, bloated, and outdated — and nobody seemed to care.

Deciding to Save the World

This is the context in which I decided to save the world. I wasn't as young and stupid as you might think — I didn't *believe* I was going to save the world. I just *decided* to. The world is big, and I was small. I knew that, in all likelihood, I'd struggle ineffectually for decades and achieve only a bitter, cynical adulthood.

But the vast majority of my peers hadn't made it as far as I had. Even though a few were sympathetic, there was simply no way we could change things. It was outside of our control.

The adults were worse. They smiled, they nodded, they commended my critical thinking skills. Then they went back to what they were doing. A few of them took the time to inform me that it's great to want to change the world and all, but eventually I'd realize that the best way to do that was to settle down and be a teacher, or run a church, or just be kind to others.

I wasn't surprised. I already knew it was rare for people to actually try and fix things.

I had youthful idealism, I had big ambitions, but I knew full well that I didn't actually have a chance. I *knew* that I wouldn't be able to single-handedly redesign the social contract, but I also knew that if everyone who made it as far as I did gave up just because changing the world is impossible, then the world would never change.

If everybody was cowed by the simple fact that they can't succeed, then that one-in-a-million person who *can* succeed would never take their shot.

So I was sure as hell going to take mine.

Broadening Scope

Mere impossibility was never a hurdle: The Phantom Tollbooth saw to that at a young age. When grownups say you can't do something, what they mean is that *they* can't do it. I spent time devising strategies to get leverage and push governments out of their stagnant state and into something capable of growth.

In 2005, a teacher to whom I'd ranted introduced me to another important book: Ishmael. It wasn't the ideas that stuck with me — I disagreed with a few at the time, and I now disagree with most. No, what this book gave me was *scope*. This author, too, wished to save the world, and the breadth of his ideas exceeded my own. This book gave me no answers, but it gave me better questions

Why merely hone the government, instead of redesigning it altogether?

More importantly, *What sort of world are you aiming for?*

"So you want to be an idealist?", the book asked. "Very well, but *what is your ideal?*"

I refocused, looking to fully define the ideals I strove for in a human social system. I knew I wouldn't be able to institute any solution directly, but I also knew that pushing governments would be much easier if I had something to push them *towards*.

After all, the Communist Manifesto changed the world, once.

This became my new goal: distill an ideal social structure for humans. The problem was insurmountable, of course, but this was hardly a deterrence. I was bright enough to understand truisms like "no one system will work for everybody" and "you're not perfect enough to get this right", but these were no trouble. I didn't need to directly specify an ideal social structure: a meta-structure, an imperfect system that ratchets towards perfection, a system that is optimal in the limit, would be fine by me.

From my vantage point, old ideas like communism and democracy soon seemed laughable. Interesting ideas in their time, perhaps, but obviously doomed to failure. It's easy to build a utopia when you imagine that people will set aside their greed and overcome their apathy. But those aren't systems for *people*: People are greedy, and people are apathetic. I wanted something that worked — nay, thrived — when populated by actual humans, with all their flaws.

I devoted time and effort to research and study. This was dangerous, as there was no feedback loop. As soon as I stepped beyond the achievements of history, there was no way to actually test anything I came up with. Many times, I settled on one idea for a few months, mulling it over, declaring it perfect. Time and again, I later found a fatal flaw, a piece of faulty reasoning, and the whole thing came tumbling down. After many cycles, I noticed that the flaws were usually visible in advance. I became cognizant of the fact that I'd been glossing over them, ignoring them, explaining them away.

I learned not to trust my own decrees of perfection. I started monitoring my thought processes very closely. I learned to notice the little ghosts of doubt, to address them earlier and more thoroughly. (I became a staunch atheist, unsurprisingly.) This was, perhaps, the beginning of my rationalist training. Unfortunately, it was all self-directed. Somehow, it never occurred to me to read literature on how to think better. I didn't have much trust in psychological literature, anyway, and I was arrogant.

Communication Failures

It was during this period that I explicitly decided not to pursue math. I reasoned that in order to actually save the world, I'd need to focus on charisma, political connections, and a solid understanding of the machinery underlying the world's major governments. Upon graduating high school, I decided to go to a college in Washington D.C. and study political science. I double majored in Computer Science as a fallback plan, a way to actually make money as needed (and because I loved it).

I went into my Poly Sci degree expecting to learn about the mechanics of society. Amusingly enough, I didn't know that "Economics" was a field. We didn't have any econ classes in my tiny high school, and nobody had seen fit to tell me about it. I expected "Political Science" to teach me the workings of nations *including* the world economy, but quickly realized that it's about the actual *politicians*, the social peacocking, the façades. Fortunately, a required Intro to Econ class soon remedied the situation, and I quickly changed my major to Economics.

My ideas experienced significant refinement as I received formal training. Unfortunately, nobody would listen to them.

It's not that they were dismissed as childish idealism: I had graduated to larger problems. I'd been thinking long and hard about the problem for a few years, and I'd had some interesting insights. But when I tried to explain them to people, almost everyone had immediate adverse reactions.

I anticipated criticism, and relished the prospect. My ideas were in desperate need of an outside challenger. But the reactions of others were far worse than I anticipated.

Nobody found flaws in my logic. Nobody challenged my bold claims. Instead, they simply failed to understand. They got stuck three or four points before the interesting points, and could go no further. I learned that most people don't understand basic economics or game theory. Many others were entrenched in bluegreensmanship and reflexively treated my suggestions as attacks. Aspiring politicians balked at the claim that Democracy, while perhaps an important step in our cultural evolution, can't possibly be the end of the line. Still others insisted that it's useless to discuss ideals, because they can never be achieved.

In short, I found myself on the far side of a wide inferential gap.

I learned that many people, after falling into the gap, were incapable of climbing out, no matter how slowly I walked them through the intervening steps. They had already passed judgement on the conclusion, and rejected my attempts to root out their misconceptions, becoming impatient before actually listening. I grew very cautious with who I shared my ideas with, worrying that exposing them too quickly or in the wrong fashion would be a permanent setback.

I had a small few friends who knew enough economics and other subjects to follow along and who wouldn't discard uncouth ideas outright. I began to value these people highly, as they were among the few who could actually put pressure on me, expose flaws in my reasoning, and help me come up with solutions.

Eventually, I had a few insights that I've yet to find in the literature, a few ideas that I still actually believe are important. You'll excuse me if I don't mention them here: there is a lot of inferential distance. Perhaps one day I'll write a sequence.

Even then, I could see no easy path to public support. Most people lacked the knowledge to understand my claims without effort, and lacked the incentive to put in

the effort for some unproven boy.

Phase Two

Fortunately, I had other tricks up my sleeve.

I attempted three different tech startups. Two of them failed. The last was healthier, but we shut it down because the expected gains were lower than an industry salary. In the interim, I honed my programming skills and secured an industry job (I'm a software engineer at Google).

By the time I graduated, my ideas were largely refined and stable. I had settled upon a solid meta social system as an ideal to strive for, and I'm still fairly confident that it's a good one — one where the design is forced by nature at every step, one that requires no arbitrary choices, one that ratchets towards optimality. And even if the ideal was not perfect, the modern world is insane enough that even a small step towards a better-coordinated society would yield gigantic benefits.

The problem changed from one of refining ideas to one of convincing others.

It was clear that I couldn't spread my ideas by merely stating them, due to the inferential distance, so I started working on two indirect approaches in the hours after work.

The first was a book, which went back to my roots: simple, low-cost ideas for how to change the *current* system of government in small ways that could have large payoffs. The goal of this project was to shake people from the blue-green mindset, to convince them that we should stop bickering within the framework and consider modifying the framework itself. This book was meant to the be first in a series, in which I'd slowly build towards more radical suggestions.

The second project was designed to put people in a more rational frame of mind. I wanted people who could look past the labels and see the *things*, people who don't just memorize how the world works but see it as mutable, as something they can actually change. I wanted people that I could pull out of inferential gaps, in case they fell into mine.

Upon introspection, I realized that much of my ability came from a specific outlook on the world that I had at a young age. I had a knack for understanding what the teachers were *trying* to teach me, for recognizing and discarding the cruft in their statements. I saw many fellow students putting stock in historical accidents of explanation where I found it easy to grasp the underlying concepts and drop the baggage. This ability to cull the cruft is important to understanding my grand designs.

This reasoning (and a few other desires, including a perpetual fascination with math and physics) led me to create simplifience, a website that promotes such a mindset.

It never made it to the point where I was comfortable publicizing it, but that hardly matters anymore. In retrospect, it's an unfinished jumble of rationality training, math explanations, and science enthusiasm. It's important in one key respect:

As I was writing simplifience, I did a lot of research for it. During this research, I kept stumbling upon web articles on this one website that articulated what I was trying to

express, only better. That website was LessWrong, and those articles were the Sequences.

It took me an embarrassingly long time to actually pay attention. In fact, if you go to simplifience.com, you can watch as the articles grow more and more influenced by the sequences. My exposure to them was patchy, centered around ideas that I'd already had. It took me a while to realize that I should read the rest of them, that I might learn new things that extended the ideas I'd figured out on my own.

It seemed like a good way to learn how to think better, to learn from someone who had had similar insights. I didn't even consider the possibility that this author, too, had some grand agenda. The idea that Eliezer's agenda could be more pressing than my own never even crossed my mind.

At this point, you may be able to empathize with how I felt when I first realized the importance of an intelligence explosion.

Superseded

It was like getting ten years worth of wind knocked out of me.

I saw something familiar in the sequences — the winding, meticulous explanations of someone struggling to bridge an inferential gap. I recognized the need to cover subjects that looked completely tangential to the actual point, just to get people to the level where they wouldn't reject the main ideas out-of-hand. I noticed the people falling to the side, debating issues two or three steps before the actual interesting problems. It was this familiar pattern, above all else, that made me *actually* pay attention.

Everything clicked. I was already thoroughly convinced of civilizational inadequacy. I had long since concluded that there's not much that can hold a strong intelligence down. I had a sort of vague idea that an AI would seek out "good" values, but such illusions were easily dispelled — I was a moral relativist. And the stakes were as high as stakes go. Artificial intelligence was a problem more pressing than my own.

The realization shook me to my core. It wasn't even the intelligence explosion idea that scared me, it was the revelation of a fatal flaw at the foundation of my beliefs. Poorly designed governments had awoken my fear that society can't handle coordination problems, but I never — not once in nearly a *decade* — stopped to consider whether designing better social systems was actually the best way to optimize the world.

I professed a desire to save the world, but had misunderstood the playing field so badly that existential risk had never even crossed my mind. Somehow, I had missed the most important problems, and they should have been obvious. Something was *very* wrong.

It was time to halt, melt, and catch fire.

This was one of the most difficult things I've done.

I was more careful, the second time around. The Sequences shook my foundations and brought the whole tower crashing down, but what I would build in its place was by

no means a foregone conclusion.

I had been blind to all existential risks, not just AI risk, and there was a possibility that I had missed other features of the problem space as well. I was well aware of the fact that, having been introduced to AI risk by Eliezer's writings, I was biased towards his viewpoint. I didn't want to make the same mistake twice, to jump for the *second* big problem that crossed my path just because it was larger than the first. I had to start from scratch, reasoning from the beginning. I knew I must watch out for conjunction fallacies caused by nice narratives, arguments made from high stakes (Pascal's mugging), putting too much stock on inside views, and so on. I had to figure out how to *actually* save the world.

It took me a long time to deprogram, to get back to neutral. I considered carefully, accounting for my biases as best I could. I read a lot. I weighed the evidence. The process took many months.

By July of 2013, I came to agree with MIRI's conclusions.

Disclaimer

Writing it all out like this, I realize that I've failed to convey the feeling of it all. Depending upon whether you believe that I was actually able to come up with better ways to structure people, you may feel that I'm either pretty accomplished or extremely deluded. Perhaps both.

Really, though, it's neither. This raw story, which omits details from the rest of my life, paints a strange picture indeed. The intensity is distilled.

I was not a zealot, in practice. My attempts to save the world didn't bleed much into the rest of my life. I learned early on that this wasn't the sort of thing that most people enjoyed discussing, and I was wary of inferential gaps. My work was done parallel to an otherwise normal life. Only a select few people were privy to my goals, my conclusions. The whole thing often felt disconnected from reality, just some unusual hobby. The majority of my friends, if they read this, will be surprised.

There are many holes in this summary, too. It fails to capture the dark spots. It omits the feelings of uncertainty and helplessness, the cycles of guilt at being unproductive followed by lingering depression, the wavering between staunch idealism and a conviction that my goals were nothing but a comfortable fantasy. It skips over the year I burned out, writing the whole idea off, studying abroad and building myself a healthier mental state before returning and picking everything back up.

Nothing in this summary describes the constant doubt about whether I was pursuing the best path or merely the easiest one. I've failed to mention my complete failure to network and my spectacular inability to find people who would actually take me seriously. It's hard to convey the fear that I was just *pretending* I wanted to save the world, just *acting* like I was trying, because that's the narrative that I wanted. How could someone 'smart' actually fail to find powerful friends if they were really trying for *nine years*?

I claim no glory: the journey was messy, and it was poorly executed. I tell the story in part because people have asked me where my passion comes from and how I became aligned with MIRI's mission. Mostly, though, I tell the story because it feels like something I have to tell before moving on. It feels almost dishonest to try to save the

world in this new way without at least acknowledging that I walked another path, once.

The source of my passion

So to those of you wondering where my passion comes from, I answer this: it has always been there. It was a small flame, when I was young, and it was fed by a deep mistrust in society's capabilities and a strong belief that if *anyone* can matter then I had better try.

From my perspective, I've been dedicating my energy towards 'saving the world' since first I realized that the world was in need of saving. This passion was not recently kindled, it was merely redirected.

There was a burst of productivity these past few months, after I refocused my efforts. I was given a new path, and on it the analogous obstacles have already been surmounted. MIRI has already spent years promoting that rational state of mind, bridging its inferential gap, finding people who can actually work on solving the problem instead of arguing about whether there is a problem to be solved. This was invigorating, like skipping ahead ten years in terms of where I wanted to be.

Alongside that, I felt a burning need to catch up. I was late to the party, and I had been foolish for a very long time. I was terrified that I wouldn't actually be able to help — that, after all my work, the most I'd be able to do to solve the big problems was earn to give. I'd have done it, because the actual goal is to save the world, not to satisfy Nate. But the idea scared me, and the desire to keep actively working on the big problems drove me forward.

In a way, too, everything got easier — I needed only to become good at logic and decision theory, to read a bunch of math textbooks, a task that was trivially measurable and joyfully easy compared to trying to convince the entire world to embrace strange, unpolished ideas.

All these factors contributed to my recent productivity. But the passion, the fervor, the desire to optimize the future — that has been there for a long time. People sometimes ask where I get my passion from, and I find it hard to answer.

We hold the entire future of the universe in our hands. Is that not justification enough?

I learned a long time ago that most people are content to accept the way things are. Everyone wants the world to change, but most are cowed by the fact that they can't change it themselves.

But if the chance that one person can save the world is one in a million, then there had better be a million people trying.

It is this knowledge — that the world will only be saved by people who actually try to save it — that drives me.

I still have these strange ideas, this pet inferential gap that I hope to bridge one day. It still hurts, that things important to me were superseded, but they *were* superseded, and it is better to know than to remain in the dark.

When I was fourteen, I saw many horrors laid out before us: war, corruption, environmental destruction, and the silent tragedies of automobile accidents, courtroom injustices, and death by disease and aging. All around me, I saw a society that couldn't coordinate, full of people resigned to unnecessary fates.

I was told to settle for making a small difference. I resolved to do the opposite.

I made a promise to myself. I didn't promise to fix governments: that was a means to an end, a convenient solution for someone who didn't know how to look further out. I didn't promise to change the world, either: every little thing is a change, and not all changes are good. No, I promised to *save* the world.

That promise still stands.

The world sure as hell isn't going to save itself.

The mechanics of my recent productivity

A decade ago, I decided to save the world. I was fourteen, and the world certainly wasn't going to save itself.

I fumbled around for nine years; it's surprising how long one can fumble around. I somehow managed to miss the whole idea of existential risk and the whole concept of an intelligence explosion. I had plenty of other ideas in my head, and while I spent a lot of time honing them, I wasn't particularly looking for new ones.

A year ago, I finally read the LessWrong sequences. My road here was roundabout, almost comical. It took me a while to come to terms with the implications of what I'd read.

Five months ago, after resolving a few internal crises, I started donating to MIRI and studying math.

Three weeks ago, I attended the December MIRI workshop on logic, probability, and reflection. I was invited to visit for the first two days and stay longer if things went well. They did: I was able to make some meaningful contributions.

On Saturday I was invited to become a MIRI research associate.

It's been an exciting year, to say the least.

(ETA: Note that being a research associate gives me access to a number of MIRI resources, but is not a full time position. I will be doing FAI research, but it will be done outside of work. I will be retaining my day job and continuing to donate.)

(ETA: As of 1 April 2014, I am a full-time researcher at MIRI.)

(ETA: As of 1 June 2015, I am now the executive director of MIRI.)

To commemorate the occasion — and because a few people have expressed interest in my efforts — I'll be writing a series of posts about my experience, about what I did and how I did it. This is the first post in the series.

First and foremost, know that I am not done with my aggressive autodidacting. I have a long way to go yet before I'm anywhere near as productive as others who do research with MIRI. I find myself at a checkpoint of sorts, collecting my thoughts in the wake of my first workshop, but next week I will be back to business.

One goal of this post is to give you a feel for how much effort is required to become good at MIRI-relevant mathematics in a short time, and perhaps inspire others to follow my path. It was difficult, but not as difficult as you might think.

Another goal is to provide data for fellow autodidacts. At the least I can provide you with an anchor point, a single datum about how much effort is required to learn at this pace. As always, remember that I am only one person and that what worked for me may not work for you.

In order to understand what I achieved it's important to know where I started from. Thus, allow me to briefly discuss my relevant prior experience.

Background

I was born in 1989. I have bachelor's degrees of science in both computer science and economics. I started programming TI-83 calculators in late 2002. I've been programming professionally since 2008. I currently work for Google and live in Seattle.

In high school I had a knack for math. I was placed two years ahead of my classmates. I aced some AP tests, I won some regional math competitions, nothing much came of it. I explicitly decided not to pursue mathematics: I reasoned that in order to save the world I would need charisma, knowledge of how the world economy works, and a reliable source of cash. This (and my love of programming) drove my choice of majors.

During college I soaked up computer science like a sponge. (Economics, too, but that's not as relevant here.) I came out of college with a strong understanding of the foundations of computing: algorithms, data structures, discrete math, etcetera. I cultivated a love for information theory. Outside of the computer science department I took two math classes: multi variable calculus and real analysis.

I was careful not to let schooling get in the way of my education. On my own time I learned Haskell in 2008 and started flirting with type theory and category theory. I read *Gödel, Escher, Bach* early in 2011.

This should paint a rough picture of my background: I never explicitly studied mathematical logic, but my interests never strayed too far from it. While I didn't have much formal training in this particular subject area, I certainly wasn't starting from a blank slate.

Accomplishments

In broad strokes, I'm writing this because I was able to learn a lot very quickly. In the space of eighteen weeks I went from being a professional programmer to helping Benja discover Fallenstein's Monster, a result concerning tiling agents (in the field of mathematical logic).

I studied math at a fervent pace from August 11th to December 12th and gained enough knowledge to contribute at a MIRI workshop. In that timeframe I read seven textbooks, five of which I finished:

1. Heuristics and Biases
2. Cognitive Science
3. Basic Category Theory for Computer Scientists
4. Naïve Set Theory
5. Model Theory (first half)
6. Computability and Logic
7. The Logic of Provability (first half, unreviewed)

In retrospect, the first two were not particularly relevant to MIRI's current research. Regardless, *Heuristics and Biases* was quite useful on a personal level.

I also studied a number of MIRI research papers, two of which I summarized:

- The Probabilistic Logic paper
- The Tiling Agents paper

I made use of a number of other minor resources as well, mostly papers found via web search. I successfully signaled my competence and my drive to the right people. While this played a part in my success, it is not the focus of this post.

I estimate my total study time to be slightly less than 500 hours. I achieved high retention and validated my understanding against other participants of the December workshop. I did this without seriously impacting my job or my social life. I retained enough spare time to participate in NaNoWriMo during November.

In sum, I achieved a high level of productivity for an extended period. In the remainder of this post I'll discuss the mechanics of how I did this: my study schedule, my study techniques, and so on. The psychological aspects — where I found my drive, how I avoid akrasia — will be covered in later posts.

Schedule

I estimate I studied 30-40 hours per week except in November, when I studied 5-15 hours per week. On average, I studied six days a week.

On the normal weekday I studied for an hour and a half in the morning, a half hour during lunch, and three to four hours in the evening. On the average weekend day I studied 8 to 12 hours on and off throughout the day.

Believe it or not, I didn't have to alter my schedule much to achieve this pace. I've been following roughly the same schedule for a number of years: I aim to spend one evening per workweek and one day per weekend on social endeavors and the rest of my time toying with something interesting. This is a loose target, I don't sweat deviations.

There were some changes to my routines, but they were minimal:

- I have many side projects, most were dropped as studying took precedence.
- The number of weeknights I took off per week fell from a little more than one to a little less than one.
- Before this endeavor I traveled for leisure about once every two months. In the past five months I traveled for leisure once.

While my studying did not affect my schedule much, it *definitely* affected my pacing. Don't get me wrong; this sprint was not easy. I suspended many other projects and drastically increased my intensity and my pace. I spent roughly the same amount of time per day studying as I used to spend on side projects, but there is a *vast* difference between spending three hours casually tinkering on open source code and spending three hours learning logic as fast as possible.

The point here is that aggressive autodidacting certainly takes quite a bit of time and effort, but it need not be all consuming: you can do this sort of thing and maintain a social life.

Study Technique

My methods were simple: read textbooks, do exercises, rephrase and write down the hard parts.

I had a number of techniques for handling difficult exercises. First, I'd put them aside and come back to them later. If that failed, I'd restate the problem (and all relevant material) in my own words. If this didn't work, it at least helped me identify the point of confusion, which set me up for a question math.stackexchange.com.

I wasn't above skipping exercises when I was convinced that the exercise was tedious and that I know the underlying material.

This sounds cleaner than it was: I made a lot of stupid mistakes and experienced my fair share of frustration. For more details on my study methods refer to On Learning Difficult Things, a post I wrote while in the midst of my struggles.

Upon finishing a book, I would immediately start the next one. Concurrently, I would start writing a review of the book I'd finished. I generally wrote the first draft of my book reviews on the Sunday after completing the book, alternating between studying the new and summarizing the old. On subsequent weekdays I'd edit in the morning and study in the evening until I was ready to post my review.

It's worth noting that summarizing content, especially the research papers, went a long way towards solidifying my knowledge and ensuring that I wasn't glossing over anything.

Impact on Social Life

The impact on my social life was minimal. I decreased contact with some periphery friend groups but maintained healthy relationships within my core circles. That I was able to do this is due in part to my circumstances:

- I live with two close friends. This meant that social contact was never out of reach. Even when spending an entire day sequestered in my room pouring over a textbook I was able to maintain a small amount of social interaction. If ever I had a spare hour and a thirst for company, I found it readily available.
- My primary partner was, up until early 2014, going to school full time while holding down a full time job. Thus, her schedule was more restrictive than my own and we had been working around it for some time. Our relationship was not further constrained by my efforts.
- My core friend groups knew and respected what I was doing. I was more tense and exhausted than usual, but I had warned my friends to expect this and no friendships suffered as a result.

Impact on Work Life

The additional cognitive load did have an impact on my day job. I had less focus and willpower to dedicate to work. Fortunately, I was exceeding expectations before this endeavor. During this sprint, with my cognitive reserves significantly depleted, I had

to settle for merely meeting expectations. My performance at work was not poor, by any means: rather, it fell from "exemplary" to "good".

I'd rather not settle for merely good performance at work for any extended period of time. Going forward, I'll be reducing my pace somewhat, in large part to ensure that I can dedicate appropriate resources to my day job.

Mental Health

It's not like I was working from dawn till dusk every day. There was ample time for other activities: I had a few hours of downtime on the average day to read books or surf the web. I participated in a biweekly Pathfinder campaign and spent the occasional Sunday playing Twilight Imperium. In September I went camping in the Olympic mountain range. I spent four days in October visiting friends in Cape Cod. I spent a day in December hiking to some hot springs. I entertained guests, went to birthday parties, and so on. There were ample opportunities to get away from math textbooks.

Most important of all, I had friends I could call on when I needed a mental health day. I could rely on them to find time where we could just sit around, play with LEGO bricks, and shoot the breeze. This went a long way towards keeping me sane.

All that said, this stint was rough. I experienced far more stress than my norm. I lost a little weight and twice caught myself grinding my teeth in my sleep (a new experience). There were days that I became mentally exhausted, growing obstinate and stubborn as if sleep- or food-deprived. This tended to happen immediately before planned breaks in the routine, as if my mind was rebelling when it thought it could get away with it.

The stress was manageable, but built up over time. It's hard to tell whether the stress was cumulative or whether the increase was due to circumstance. Doing NaNoWriMo in November while continuing studying didn't particularly help matters. The weeks leading up to the workshop were particularly stressful due to a lack of information: I worried that I would not know nearly enough to be useful, that I would make a fool of myself, and so on. So while the stress surely mounted as time wore on, I can't tell how much of that was cumulative versus circumstantial.

I tentatively believe that someone could sustain my pace for significantly longer than I did, so long as they were willing to live with the strain. I don't plan to test this myself: I'll be slowing down both to improve performance at work and to reduce my general stress levels. Five months of fervent studying is no walk in the park.

Advice

So you want to follow in my footsteps? Awesome. I commend your enthusiasm. My next post will delve into my mindset and a few of the quirks of my behavior that helped me be productive. For now, I will leave you with this advice:

- There is no magic to it. If you study the right material, do the exercises, and write what you've learned in your own words, then you can indeed learn MIRI-relevant math in a reasonable amount of time.

- Learning fast does not need to dominate your life. There can be time for social activities and even significant side projects. You will have to work really hard, but that work does not have to consume your life.
- If you're going to do something like this, let people know what you're doing. This is much easier if you have people you can turn to for support who don't mind you being extra snappy, people who can drag you away for a day every week or two. Also, stating your goals publicly helps to stop you from giving up.

The difficult part is making a commitment and sticking to it. Akrasia is a formidable enemy, here. If you can avoid it, the actual autodidacting is not overly difficult.

As for specific advice, if your background is similar to mine then I recommend reading *Naïve Set Theory, Computability and Logic,* and the first two chapters of *Model Theory* in that order, these will get you off to a good start. Feel free to PM me if you get stuck or if you want more recommendations.

Following posts will cover the other sides of my experience: how I got interested in this field, where I draw my motivation from, and the dark arts that I use to maintain productivity. In the meantime, questions are welcome.

Dangers of steelmanning / principle of charity

As far as I can tell, most people around these parts consider the principle of charity and its super saiyan form, steelmanning, to be Very Good Rationalist Virtues. I basically agree and I in fact operate under these principles more or less automatically now. HOWEVER, no matter how good the rule is, there are always exceptions, which I have found myself increasingly concerned about.

This blog post that I found in the responses to Yvain's anti-reactionary FAQ argues that even though the ancient Romans had welfare, this policy was motivated not for concern for the poor or for a desire for equality like our modern welfare policies, but instead "the Roman dole was wrapped up in discourses about a) the might and wealth of Rome and b) goddess worship... The dole was there because it made the emperor more popular and demonstrated the wealth of Rome to the people. What's more, the dole was personified as Annona, a goddess to be worshiped and thanked."

So let's assume this guy is right, and imagine that an ancient Roman travels through time to the present day. He reads an article by some progressive arguing (using the rationale one would typically use) that Obama should increase unemployment benefits. "This makes no sense," the Roman thinks to himself. "Why would you give money to someone who doesn't work for it? Why would you reward lack of virtue? Also, what's this about equality? Isn't it right that an upper class exists to rule over a lower class?" Etc.

But fortunately, between when he hopped out of the time machine and when he found this article, a rationalist found him and explained to him steelmanning and the principle of charity. "Ah, yes," he thinks. "Now I remember what the rationalist said. I was not being so charitable. I now realize that this position kind of makes sense, if you read between the lines. Giving more unemployment benefits *would*, now that I think about it, demonstrate the power of America to the people, and certainly Annona would approve. I don't know why whoever wrote this article didn't just come out and say that, though. Maybe they were confused".

Hopefully you can see what I'm getting at. When you regularly use the principle of charity and steelmanning, you run the risk of:

1. Sticking rigidly to a certain worldview/paradigm/established belief set, even as you find yourself willing to consider more and more concrete propositions. The Roman would have done better to really read what the modern progressive's logic was, think about it, and try to see where he was coming from than to automatically filter it through his own worldview. If he consistently does this he will never find himself considering alternative ways of seeing the world that might be better.

2. Falsely developing the sense that your worldview/paradigm/established belief set is more popular than it is. Pretty much no one today holds the same values that an ancient Roman does, but if the Roman goes around being charitable all the time then he will probably see his own beliefs reflected back at him a fair amount.

3. Taking arguments more seriously than you possibly should. I feel like I see all the time on rationalist communities people say stuff like "this argument by A sort of makes sense, you just need to frame it in objective, consequentialist terms like blah

blah blah blah blah" and then follow with what looks to me like a completely original thought that I've never seen before. But why didn't A just frame her argument in objective, consequentialist terms? Do we assume that what she wrote was sort of a telephone-game approximation of what was originally a highly logical consequentialist argument? If so where can I find that argument? And if not, why are we assuming that A is a crypto-consequentialist when she probably isn't? And if we're sure that objective, consequentialist logic is The Way To Go, then shouldn't we be very skeptical of arguments that seem like their basis is in some other reasoning system entirely?

4. Just having a poor model of people's beliefs in general, which could lead to problems.

Hopefully this made sense, and I'm sorry if this is something that's been pointed out before.

Rationalists Are Less Credulous But Better At Taking Ideas Seriously

Consider the following commonly-made argument: cryonics is unlikely to work. Trained rationalists are signed up for cryonics at rates much greater than the general population. Therefore, rationalists must be pretty gullible people, and their claims to be good at evaluating evidence must be exaggerations at best.

This argument is wrong, and we can prove it using data from the last two Less Wrong surveys.

The question at hand is whether rationalist training - represented here by extensive familiarity with Less Wrong material - makes people more likely to believe in cryonics.

We investigate with a cross-sectional study, looking at proto-rationalists versus experienced rationalists. Define proto-rationalists as those respondents to the Less Wrong survey who indicate they have been in the community for less than six months and have zero karma (usually indicative of never having posted a comment). And define experienced rationalists as those respondents to the Less Wrong survey who indicate they have been in the community for over two years and have >1000 karma (usually indicative of having written many well-received posts).

By these definitions, there are 93 proto-rationalists, who have been in the community an average of 1.3 months, and 134 experienced rationalists, who have been in the community an average of 4.5 years. Proto-rationalists generally have not read any rationality training material - only 20/93 had read even one-quarter of the Less Wrong Sequences. Experienced rationalists are, well, more experienced: two-thirds of them have read pretty much all the Sequence material.

Proto-rationalists thought that, on average, there was a 21% chance of an average cryonically frozen person being revived in the future. Experienced rationalists thought that, on average, there was a 15% chance of same. The difference was marginally significant ($p < 0.1$).

Marginal significance is a copout, but this isn't our only data source. Last year, using the same definitions, proto-rationalists assigned a 15% probability to cryonics working, and experienced rationalists assigned a 12% chance. We see the same pattern.

So experienced rationalists are consistently less likely to believe in cryonics than proto-rationalists, and rationalist training probably makes you less likely to believe cryonics will work.

On the other hand, 0% of proto-rationalists had signed up for cryonics compared to 13% of experienced rationalists. 48% of proto-rationalists rejected the idea of signing up for cryonics entirely, compared to only 25% of experienced rationalists. So although rationalists are less likely to believe cryonics will work, they are much more likely to sign up for it. Last year's survey shows the same pattern.

This is not necessarily surprising. It only indicates that experienced rationalists and proto-rationalists treat their beliefs in different ways. Proto-rationalists form a belief,

play with it in their heads, and then do whatever they were going to do anyway - usually some variant on what everyone else does. Experienced rationalists form a belief, examine the consequences, and then act strategically to get what they want.

Imagine a lottery run by an incompetent official who accidentally sets it up so that the average payoff is far more than the average ticket price. For example, maybe the lottery sells only ten $1 tickets, but the jackpot is $1 million, so that each $1 ticket gives you a 10% chance of winning $1 million.

Goofus hears about the lottery and realizes that his expected gain from playing the lottery is $99,999. "Huh," he says, "the numbers say I could actually *win* money by playing this lottery. What an interesting mathematical curiosity!" Then he goes off and does something else, since everyone knows playing the lottery is what stupid people do.

Gallant hears about the lottery, performs the same calculation, and buys up all ten tickets.

The relevant difference between Goofus and Gallant is not skill at estimating the chances of winning the lottery. We can even change the problem so that Gallant is more aware of the unlikelihood of winning than Goofus - perhaps Goofus mistakenly believes there are only five tickets, and so Gallant's superior knowledge tells him that winning the lottery is even more unlikely than Goofus thinks. Gallant will still play, and Goofus will still pass.

The relevant difference is that Gallant knows how to take ideas seriously.

Taking ideas seriously isn't always smart. If you're the sort of person who falls for proofs that 1 = 2 , then refusing to take ideas seriously is a good way to avoid ending up actually believing that 1 = 2, and a generally excellent life choice.

On the other hand, progress depends on someone somewhere taking a new idea seriously, so it's nice to have people who can do that too. Helping people learn this skill and when to apply it is one goal of the rationalist movement.

In this case it seems to have been successful. Proto-rationalists think there is a 21% chance of a new technology making them immortal - surely an outcome as desirable as any lottery jackpot - consider it an interesting curiosity, and go do something else because only weirdos sign up for cryonics.

Experienced rationalists think there is a lower chance of cryonics working, but some of them decide that even a pretty low chance of immortality sounds pretty good, and act strategically on this belief.

This is not to either attack or defend the policy of assigning a non-negligible probability to cryonics working. This is meant to show only that the difference in cryonics status between proto-rationalists and experienced rationalists is based on meta-level cognitive skills in the latter whose desirability is orthogonal to the object-level question about cryonics.

(an earlier version of this article was posted on my blog last year; I have moved it here now that I have replicated the results with a second survey)

Handshakes, Hi, and What's New: What's Going On With Small Talk?

This is an attempt to explicitly model what's going on in some small talk conversations. My hope is that at least one of these things will happen:

- There is a substantial flaw or missing element to my model that someone will point out.
- Many readers, who are bad at small talk because they don't see the point, will get better at it as a result of acquiring understanding.

Handshakes

I had some recent conversational failures online, that went roughly like this:

> “Hey.”
> “Hey.”
> “How are you?”
> The end.

At first I got upset at the implicit rudeness of my conversation partner walking away and ignoring the question. But then I decided to get curious instead and posted a sample exchange (names omitted) on Facebook with a request for feedback. Unsurprisingly I learned more this way.

Some kind friends helped me troubleshoot the exchange, and in the process of figuring out how online conversation differs from in-person conversation, I realized what these things do in live conversation. They act as a kind of implicit communication protocol by which two parties negotiate how much interaction they’re willing to have.

Consider this live conversation:

> “Hi.”
> “Hi.”
> The end.

No mystery here. Two people acknowledged one another’s physical presence, and then the interaction ended. This is bare-bones maintenance of your status as persons who can relate to one another socially. There is no intimacy, but at least there is acknowledgement of someone else’s existence. A day with “Hi” alone is less lonely than a day without it.

> “Hi.”
> “Hi, how’s it going?”
> “Can’t complain. And you?”
> “Life.”

This exchange establishes the parties as mutually sympathetic – the kind of people who would ask about each other’s emotional state – but still doesn’t get to real intimacy. It is basically just a drawn-out version of the example with just “Hi”. The exact character of the third and fourth line don’t matter much, as there is no real

content. For this reason, it isn't particularly rude to leave the question totally unanswered if you're already rounding a corner – but if you're in each other's company for a longer period of time, you're supposed to give at least a pro forma answer.

This kind of thing drives crazy the kind of people who actually want to know how someone is, because people often assume that the question is meant insincerely. I'm one of the people driven crazy. But this kind of mutual "bidding up" is important because sometimes people don't want to have a conversation, and if you just launch into your complaint or story or whatever it is you may end up inadvertently cornering someone who doesn't feel like listening to it.

You could ask them explicitly, but people sometimes feel uncomfortable turning down that kind of request. So the way to open a substantive topic of conversation is to leave a hint and let the other person decide whether to pick it up. So here are some examples of leaving a hint:

"Hi."
"Hi."
"Anything interesting this weekend?"
"Oh, did a few errands, caught up on some reading. See you later."

This is a way to indicate interest in more than just a "Fine, how are you?" response. What happened here is that one party asked about the weekend, hoping to elicit specific information to generate a conversation. The other politely technically answered the question without any real information, declining the opportunity to talk about their life.

"Hi."
"Hi."
"Anything interesting happen over the weekend?"
"Oh, did a few errands, caught up on some reading."
"Ugh, I was going to go to a game, but my basement flooded and I had to take care of that instead."
"That's tough."
"Yeah."
"See you around."

Here, the person who first asked about the weekend didn't get an engaged response, but got enough of a pro forma response to provide cover for an otherwise out of context complaint and bid for sympathy. The other person offered perfunctory sympathy, and ended the conversation.

Here's a way for the recipient of a "How are you?" to make a bid for more conversation:

"Hi."
"Hi."
"How are you?"
"Oh, my basement flooded over the weekend."
"That's tough."
"Yeah."
"See you around."

So the person with the flooded basement provided a socially-appropriate snippet of information – enough to be a recognizable bid for sympathy, but little enough not to force the other person to choose between listening to a long complaint or rudely cutting off the conversation.

Here's what it looks like if the other person accepts the bid:

"Hi."
"Hi."
"How are you?"
"Oh, my basement flooded over the weekend."
"Wow, that's tough. Is the upstairs okay?"
"Yeah, but it's a finished basement so I'm going to have to get a bunch of it redone because of water damage."
"Ooh, that's tough. Hey, if you need a contractor, I had a good experience with mine when I had my kitchen done."
"Thanks, that would be a big help, can you email me their contact info?"

By asking a specific follow-up question the other person indicated that they wanted to hear more about the problem – which gave the person with the flooded basement permission not just to answer the question directly, but to volunteer additional information / complaints.

You can do the same thing with happy events, of course:

"Hi."
"Hi."
"How are you?"
"I'm getting excited for my big California vacation."
"Oh really, where are you going?"
"We're flying out to Los Angeles, and then we're going to spend a few days there but then drive up to San Francisco, spend a day or two in town, then go hiking in the area."
"Cool. I used to live in LA, let me know if you need any recommendations."
"Thanks, I'll come by after lunch?"

So what went wrong online? Here's the conversation again so you don't have to scroll back up:

"Hey."
"Hey."
"How are you?"
The end.

Online, there are no external circumstances that demand a "Hi," such as passing someone (especially someone you know) in the hallway or getting into an elevator.

If you import in-person conversational norms, the "Hi" is redundant – but instead online it can function as a query as to whether the other person is actually "present" and available for conversation. (You don't want to start launching into a conversation just because someone's status reads "available" only to find out they're in the middle of something else and don't have time to read what you wrote.)

Let's say you've mutually said "Hi." If you were conversing in person, the next thing to do would be to query for a basic status update, asking something like, "How are

you?”. But “Hi” already did the work of “How are you?”. Somehow the norm of “How are you?” being a mostly insincere query doesn’t get erased, even though “Hi” does its work – so some people think you’re being bizarrely redundant. Others might actually tell you how they are.

To be safe, it’s best to open with a short question apropos to what you want to talk about – or, since it’s costless online and serves the same function as “Hi”, just start with “How are you?” as your opener.

What’s New?

I recently had occasion to explain to someone how to respond when someone asks “what’s new?”, and in the process, ended up explaining some stuff I hadn’t realized until the moment I tried to explain it. So I figured this might be a high-value thing to explain to others here on the blog.

Of course, sometimes “what’s new?” is just part of a passing handshake with no content – I covered that in the first section. But if you’re already in a context where you know you’re going to be having a conversation, you’re supposed to answer the question, otherwise you get conversations like this:

> “Hi.”
> “Hi.”
> “What’s new?”
> “Not much. How about you?”
> “Can’t complain.”
> Awkward silence.

So I’m talking about cases where you actually have to answer the question.

The problem is that some people, when asked “What’s New?”, will try to think about when they last met the person asking, and all the events in their life since then, sorted from most to least momentous. This is understandably an overwhelming task.

The trick to responding correctly is to think of your conversational partner’s likely motives for asking. They are very unlikely to want a complete list. Nor do they necessarily want to know the thing in your life that happened that’s objectively most notable. Think about it – when’s the last time you wanted to know those things?

Instead, what’s most likely the case is that they want to have a conversation about a topic you are comfortable with, are interested in, and have something to say about. “What’s New?” is an offer they are making, to let you pick the life event you most feel like discussing at that time. So for example, if the dog is sick but you’d rather talk about a new book you’re reading, you get to talk about the book and you can completely fail to mention the dog. You’re not lying, you’re answering the question as intended.

Cross-posted on my personal blog.

Deregulating Distraction, Moving Towards the Goal, and Level Hopping

This is the third post in a series discussing my recent bout of productivity. Within, I discuss two techniques I use to avoid akrasia and one technique I use to be especially productive.

Deregulating Distraction

I like to pretend that I have higher-than-normal willpower, because my ability to Get Things Done seems to be somewhat above average. In fact, this is not the case. I'm not good at fighting akrasia. I merely have a knack for avoiding it.

When I was young, my parents were very good at convincing me to manage my money. They gave me an allowance, perhaps a dollar a week. When we would go to the store, I'd get excited about some trite toy and ask my parents whether I could buy it.

Their answers were similar. My mother would crouch down, put a hand on my shoulder, and say "Of course you can. But before you do, think carefully about how much you will enjoy this after you've bought it, and what other things you would be able to buy if instead you saved up."

My father was a bit more direct. He'd just shrug and say "It's your money", with the barest hint of derision.

I rarely spent my allowance.

I now use a similar technique when dealing with distractions.

(It's worth noting that it's always been very easy to put me into far mode, perhaps in part because I decided at a very young age that I wasn't going to die.)

As Kaj Sotala and a few others noted, assigning guilt to non-productive tasks is not especially healthy. Nor is it, in my experience, sustainable. In a few different cases, I experienced scenarios where I wanted to do something but couldn't will myself to do it. I suffered ego depletion and hit a vicious cycle of unproductivity and depression. I never fell completely into the self-hate death spiral, but I flirted around at the edges. It became clear that I needed a new strategy.

To break the cycle, I decided to stop fighting myself.

The world is full of distractions, and I have plenty of vices. I am just as susceptible as anyone to binging on TV shows or video games or book series. Instead of trying (and often failing) to stop myself from indulging, I decided to allow myself to indulge whenever I really wanted to.

"It's your time", I told myself.

This changed the game entirely. I no longer willed myself to avoid temptation: I weighed temptations alongside my other options, took their pros and cons into

account, and made an informed decision. Did I *need* to distract myself? Sometimes, the answer was yes.

Knowing that I could no longer trust myself to bail me out if I got addicted to new media, I took special care in removing as many distractions as I could from my environment. Because I'd resolved not to spend willpower to cancel addictions, I became much more cautious at the point of entry. These days, I ignore recommendations about new TV shows and books, preferring not even to learn the premises, thus dodging the temptation entirely.

By allowing distractions a place in my mental calculus I allowed myself to choose between them with more care: I am able to watch movies instead of TV shows, to read standalone books instead of entire series.

I know full well that my resolution against spending willpower against myself means that once I get addicted to something, it has to run its full course before I can be productive again. This is a nuclear option: because I know that I *won't* stop, I am *very* leery of lengthy media. I avoid open-ended addictions (ongoing online games, chemical addictions, etc.) like the plague.

I refer to this strategy as "playing chicken against myself": because I know that I'll let long addictions run their course, I seldom have to.

From another perspective, you could say that I deregulated a black market on distractions: By lifting the mental ban on entertainment, I was able to price it accurately and weigh the tradeoffs. If there is a new book I want to read, the answer is not an outright and unenforcible "No". Rather, it's "can we afford to be underproductive for the next few days?". And when the answer is negative, it's significantly easier for me to postpone gratification than to resist the temptation entirely. The end result is that I have much more control over when I indulge in escapism.

Finally, I've found that this *feels* a lot better than feeling guilty about being unproductive. It's a healthier state of mind, and it's led to a general increase in happiness.

Moving Towards the Goal

My teachers used to tell my parents that I have two modes of operation: I either put in the minimum possible effort or I blow expectations completely out of the water. They claimed I have no middle ground.

This isn't quite accurate. The truth is, I *always* put in the minimum effort. Anything else would be wasted motion. The discrepancy they observed was not due to some whim of passion, it was an artifact of how our incentives were misaligned.

In school I was incentivized to ace classes with minimal work. I was very good at obeying the letter of the law while blatantly flouting the spirit, and I had a knack for knowing *exactly* how far I could push my luck. My teachers had… polarized opinions of me, to say the least. I was an arrogant kid.

Yet when my schoolwork happened to align with some personal goal — mastering a new technique, figuring out new secrets of the universe — then I was relentless,

shattering expectations with apparent ease. A number of my teachers took it upon themselves to press upon me just how much I could do if I actually *applied* myself. I didn't bother correcting them. If they weren't going to invent a grade higher than 'A', why should I waste my efforts in the classroom? I had better things to do.

Like I said, I was an arrogant kid.

This experience in school had two important repercussions. First, it taught me to seek out the gap between the *intended* rules and the *actual* rules. I developed a knack for it, and this has served me well in many walks of life. Noticing the space between what you meant and what you said is a fundamental skill for programmers. Math is a tool designed to narrow such gaps. Logical incompleteness theorems are statements about the gap between what logic *can* say and what mathematicians *want* to say.

Secondly, and more relevant to this post, school helped me make explicit the virtue of putting in the minimum possible effort. Authority figures parroted the value of hard work, but that's only half the story. You should *always* be putting forth the least amount of effort that it takes to achieve your goals. That's not to say that you should never do hard work: in many situations, the easiest way to achieve your goals is to do things right the first time. I'm not condoning shoddy work, either: if quality is part of your goal then you'd best do things correctly. If you're trying to signal competence, then by all means, put in extra effort. But you should *never* expend extra effort just for effort's sake.

This leads us to my second trick for avoiding akrasia: I am not Trying Really Hard. People who are Trying Really Hard give themselves rewards for progress or punishments for failure. They incentivize the behavior that they want to have. They keep on deciding to continue doing what they're doing, and they engage in valiant battle against akrasia. I don't do any of that.

Instead, I simply Move Towards the Goal.

I don't will myself to study. It is not a chore, it is not something I force myself to do. That's not to say I enjoy studying, per se: it's hard work, and the reward structure is pathetic compared to programming. If I had to force or convince myself to study lots of math continously, I don't think I'd get very far.

That's not how I operate. I don't Try Really Hard. I simply Move Towards the Goal.

This is where the previous post ties in. I've mostly eliminated the guilt I feel while unproductive, but I've maintained two very important things from that era of my life:

1. In my head, long-term satisfaction is linked to productivity.
2. I have maintained habitual productivity for years.

Between these two points, I know that once I've settled on a goal, I'm going to more towards it.

This is, internally, an immutable fact, made so both by habit and by crude Pavlovian training. None of this is explicit, mind you, it's just the *nature of goals*. I can change the goal and I can drop the goal, but I can't hold the goal and *not pursue* it.

I never *decided* to study really hard. You can "decide" not to watch the next episode of that TV show only to sternly berate yourself three episodes later. My decision to study

hard was made on a lower level, it's been internalized. Acting on goals the thing that System 1 does regardless of what System 2 "decides".

System 2 controls things by *picking* the goals. It was a long and arduous process to internalize my most recent set of goals, the ones that have driven me to study hard and become a research associate and so on. It took a few months and a bit of mindhacking, and that's a story for another day. But once the goal was *chosen*, marching towards it was out of my hands.

System 2 isn't in control of *whether* I move towards the goal. Instead, it spends its time doing something it's very good at: finding the most efficient path. Minimizing effort.

I don't actively force myself to study hard. Rather, the structure of the environment is such that the shortest path to the goal requires hard studying. I merely follow that path.

Moving Towards the Goal might look a lot like Trying Really Hard from the outside. Superficially, the two are similar. On the inside, though, they feel very different. I've Tried Really Hard before, and I'm not good at it. It requires exertion of willpower and results in depletion of ego.

When I'm Moving Towards the Goal, I don't worry about whether things will be done. I've outsourced that concern to habit. Instead, mental effort is spent look for the shortest path, the easiest route. Difficult paths do not require additional willpower, because the internal narrative is not one of expending effort. If anything, a difficult path is worth extra points, because it means I'm pursuing admirable goals. Internally, I'm not Struggling Against Akrasia. I'm Finding an Efficient Route.

Don't get me wrong, studying math at high speed for five months was hard. However, I have built myself a headspace where hardness is not an obstacle to overcome but a *feature of the terrain*. I am going to march on regardless. System 2 doesn't have to spend effort convincing System 1 to move forward, because System 1 is going to move forward come hell or high water. Thus, System 2 spends its time making sure that the march is as easy as possible.

This leaves me free to try new techniques to achieve my goals more effectively, and that leads us to our final trick for the day.

Level Hopping

I started doing NaNoWriMo in 2011, and I noticed something interesting: a vast majority of winners *barely* made it to 50,000 words. The goal of NaNoWriMo is to write 50k words in a month, so I wasn't particularly surprised. However, from my interactions with others I found that a vast majority of these winners *felt* like they were pushing themselves to the limit, even though many of them were probably psychologically anchored below their actual limits. After all, in my experience, the hardest part of NaNoWriMo is *writing every day*: the most difficult part of being productive is switching contexts, once you get rolling it's not difficult to keep rolling.

It seemed clear that if the goal had been 60k, many of the same people would have eeked out a victory with similar margins and the same narrative of butting against

their limits. The natural conclusion was that I can't trust myself to feel out my own limits.

This is when I decided to start hopping to higher levels of productivity. These days, I occasionally throw wrenches into my study plans when I think I'm growing complacent.

"Those set theory and category theory books were easy", I'll say, "Let's try skipping introductory logic and going straight to model theory".

Or, "All this studying is great, but I bet I could keep it up and also do a NaNoWriMo for 75k words".

Often, this fails spectacularly. Sometimes, I *am* at or near my limits, and skipping an intro logic textbook to dive straight into Model Theory is a *really bad idea*. Other times, I find out that I actually was just hovering around an anchor point, seduced by a narrative of linear improvement.

This is not an original idea, by any means. In fact, there's a relevant Bruce Lee quote:

> There are no limits. There are plateaus, but you must not stay there, you must go beyond them. If it kills you, it kills you. A man must constantly exceed his level.

- Bruce Lee

My point, more broadly, is that this is the type of thing that occupies my mental narrative. I'm not wondering whether I will be able to convince myself to study each day. Instead, I'm gauging whether I'm reading the most effective material. I'm noticing that it won't be enough for me to just *learn* the material, I also have to *signal* that I've learned the material (and that I should start doing book reviews). I'm monitoring to see when I've grown complacent and looking for ways to keep me on my toes. This is process is doubly useful: It helps me sidestep akrasia and it also helps me become more effective.

These are my three Light Side tools:

1. I've constructed an environment in which productivity is habitual. In the absence of distractions, I trust myself to get things done.
2. I've lifted my mental ban on distractions, and trust myself to use them wisely.
3. My mental narrative is one of expending minimal effort, not one of trying to succeed: instead of worrying about whether I can continue, I worry about how to perform better.

Most of these tricks are likely familiar: I do not claim originality; this is merely an account of the methods that I use, the things that work for me. Consider this to be evidence that these techniques work for people who share my personality (which I've tried to illustrate along the way).

You now have a broad sketch of how I maintain productivity, but it may seem somewhat unstable, difficult to maintain indefinitely. The next post will detail my Dark Side tactics: tricks I use to remain unrelenting and sustain my vigorous pace, but which may make rationalists uncomfortable.

After that, I'll tell the story of a kid who decided he would save the world for reasons completely unrelated to existential risk, and how he came to align himself with MIRI's mission. This will help you understand the source of my passion, and will conclude the series.

Dr. Jubjub predicts a crisis

Dr. Jubjub: Sir, I have been running some calculations and I'm worried about the way our slithy toves are heading.

Prof. Bandersnatch: Huh? Why? The toves seem fine to me. Just look at them, gyring and gimbling in the wabe over there.

Dr. Jubjub: Yes, but there is a distinct negative trend in my data. The toves are gradually losing their slithiness.

Prof. Bandersnatch: Hmm, okay. That does sound serious. How long until it becomes a problem?

Dr. Jubjub: Well, I'd argue that it's already having negative effects but I'd say we will reach a real crisis in around 120 years.

Prof. Bandersnatch: Phew, okay, you had me worried there for a moment. But it sounds like this is actually a non-problem. We can carry on working on the important stuff – technology will bail us out here in time.

Dr. Jubjub: Sir! We already have the technology to fix the toves. The most straightforward way would be to whiffle their tulgey wood but we could also...

Prof. Bandersnatch: What?? Whiffle their tulgey wood? Do you have any idea what that would cost? And besides, people won't stand for it – slithy toves with unwhiffled tulgey wood are a part of our way of life.

Dr. Jubjub: So, when you say technology will bail us out you mean you expect a solution that will be cheap, socially acceptable and developed soon?

Prof. Bandersnatch: Of course! Prof. Jabberwock assures me the singularity will be here around tea-time on Tuesday. That is, if we roll up our sleeves and don't waste time with trivialities like your tove issue.

Maybe it's just me but I feel like I run into a lot of conversations like this around here. On any problem that won't become an absolute crisis in the next few decades, someone will take the Bandersnatch view that it will be more easily solved later (with cheaper or more socially acceptable technology) so we shouldn't work directly on it now. The way out is forward - let's step on the gas and get to the finish line before any annoying problems catch up with us.

For all I know, Bandersnatch is absolutely right. But my natural inclination is to take the Jubjub view. I think the chances of a basically business-as-usual future for the next 200 or 300 years are not epsilon. They may not be very high but they seem like they need to be seriously taken into account. Problems may prove harder than they look. Apparently promising technology may not become practical. Maybe we'll have the capacity for AI in 50 years - but need another 500 years to make it friendly. I'd prefer humanity to plan in such a way that things will gradually improve rather than gradually deteriorate, even in a slow-technology scenario.

LessWrong Help Desk - free paper downloads and more (2014)

Over the last year, VincentYu, gwern and others have provided many papers for the LessWrong community (87% success rate in 2012) through previous help desk threads. We originally intended to provide editing, research and general troubleshooting help, but article downloads are by far the most requested service.

If you're doing a LessWrong relevant project we want to help you. If you need help accessing a journal article or academic book chapter, we can get it for you. If you need some research or writing help, we can help there too.

Turnaround times for articles published in the last 20 years or so is usually less than a day. Older articles often take a couple days.

Please make new article requests in the comment section of this thread.

If you would like to help out with finding papers, please monitor this thread for requests. If you want to monitor via RSS like I do, many RSS readers will give you the comment feed if you give it the URL for this thread (or use this link directly).

If you have some special skills you want to volunteer, mention them in the comment section.

Try more things.

(Cross-posted from my personal site.)

Several months ago I began a list of "things to try," which I share at the bottom of this post. It suggests many mundane, trivial-to-medium-cost changes to lifestyle and routine. Now that I've spent some time with most of them and pursued at least as many more personal items in the same spirit, I'll suggest you do something similar. Why?

- Raise the temperature in your optimization algorithm: avoid the trap of doing too much analysis on too little data and escape local optima.
- You can think of this as a system for self-improvement; something that operates on a meta level, unlike an object-level goal or technique; something that helps you fail at almost everything but still win big.
- Variety of experience is an intrinsic pleasure to many, and it may make you feel less that time has flown as you look back on your life.
- Practice implementing small life changes, practice observing the effects of the changes, practice noticing further opportunities for changes, practice value of information calculations, and reinforce your self-image as an empiricist working to improve your life. Build small skills in the right order and you'll have better chances at bigger wins in the future.
- Advice often falls prey to the typical-mind (or typical-body) fallacy. That doesn't mean you should dismiss it out of hand. Think about not just how likely it is to work for you, but how beneficial it would be if it worked, how much it would cost to try, and how likely it is that trying it would give you enough information to change your behavior. Then just try it anyway if it's cheap enough, because you forgot to account for uncertainty in your model inputs.
- Speaking of value of information: don't ignore tweakable variables just because you don't yet have a gwern-tier tracking and evaluation apparatus for the perfect self-experiment. Sometimes you can expect consciously noticeable non-placebo effects from a successful trial. You might do better picking the low hanging fruit to gain momentum before you invest in a Zeo and a statistics textbook.
- You know what, if there's an effect, it may not even need to be non-placebo. C.f. "Lampshading," as well as the often-observed "honeymoon" period of success with new productivity systems.
- It's very tempting, especially in certain communities, to focus exclusively on shiny, counterintuitive, "rational," tech-based, hackeresque, or otherwise clever interventions and grand personal development schemes. Some of these are even good, but one suspects that some are optimized for punchiness, not effectiveness. Conversely, mundane ideas may not propagate as well, despite being potentially equally or more likely to succeed.
- If you were already convinced of all of the above, then great! I hope you have the agency to try stuff like this all the time. If not, you might find it useful, as I did, just to have a list like this available. It's one less trivial inconvenience between thinking "I should try more things" and actually trying something. I've also found that I'm more likely to notice and remember optimization opportunities now that I have a place to capture them. And having spent the time to write them down and occasionally look over them, I'm more likely to notice when I'm in a position to enact something context-dependent on the list.

I removed the terribly personal items from my list, but what remains is still somewhat tailored to my own situation and habits. These are not recommendations; they are just things that struck me as having enough potential value to try for a week or two. The list isn't not remotely comprehensive, even as far as mundane self-experiments are concerned, but it's left as an exercise to the reader to find and fill the gaps. Take this list as an example or as a starting point, and brainstorm ideas of your own in the comments. The usual recommendation applies against going overboard in domains where you're currently impulsive or unreflective.

Related posts: Boring Advice Repository, Break your habits: Be more empirical, On saying the obvious, Value of Information: Four Examples, Spend money on ergonomics, Go try things, Don't fear failure, Just try it: Quantity trumps quality, No, seriously, just try it, etc.

META

- *Before* you read the rest of this list, spend two minutes brainstorming ideas to try!
    - In what domains are you in a rut? What do you do frequently, and what are alternative ways to do it? What do others do differently? What vague dissatisfactions tickle your attention?
- Incorporate trying things from your list into your routine
- Incorporate adding things to your list into your routine
- Do something you've tried in the past (this is "try more things," not "new things")
- Attempt some value of information calculations for trying or researching items below
- Ask friends/coworkers for recommendations (or bring them in on the adventures herein)
- Create a system to reliably capture ideas before you forget them and later add them to your list (e.g. take notes on your phone)
- Learn about and implement some more rigorous self-experimentation
- Learn to break down desired new behaviors into cue, routine, reward and practice them offline

SLEEP

- Earplugs, or a change in style or brand if you already use them
- Melatonin, or vary dose and timing
- Sleep mask; an extra pillowcase as a blindfold might be sufficient
- Wear socks or slippers to bed
- Different pillows, sheets or bedding
- Side/back sleeping
- Windows open/closed
- White noise (perhaps a fan or a recording)
- Humidifier
- Air filter

- Blackout curtains (particularly if you find a sleep mask uncomfortable but like the darkness)
- Napping
- Morning vitamin D
- "Sleep-tracking" phone app (can record movement and noise, which is sometimes informative)
- Antihistamines in the case of allergies disrupting breathing
- Dream journal
- Sleep journal (notes on sleep time, quality, etc.)
- Lucid dreaming
- **BEDTIME ROUTINE**
    - Construct a nighttime ritual
    - Keep to a specific bedtime
    - Use bed for sleep only
    - Stop using a computer by T minus X hours
    - Stop working by T minus Y hours
    - Don't eat after T minus Z hours
    - Alternatively, light snack before bed
    - Don't drink after T minus V hours
    - Change lighting by T minus W hours (warm/dim lights)
    - Use flux or alternatives http://alternativeto.net/software/f46lux/
    - Stretching/breathing exercise
    - Intense exercise (most likely to be beneficial well before bedtime ritual starts; I recall reading at least three hours)
    - Further research on sleep habits
- **WAKING ROUTINE**
    - Use an alarm, or use a different sound, or place it somewhere new
        - set up difficult tasks to turn off alarm
        - use a light on a timer rather than sound
    - Don't use an alarm -- wake up to daylight or use your natural cycle
    - Practice offline (either with naps, or just getting in bed and then out again)
    - Have a 'halfway point' to getting out of bed which makes things take much less than half the effort
    - Count down from 10, intensely focusing on your plan to get out of bed when you get to 0
    - Open windows or go outside after waking (particularly if it's sunny)
    - Splash cold water on your face
    - Morning stretching , light exercise, or intense exercise routine

WORK ENVIRONMENT

- Change relative heights of chair, keyboard, monitor (can stack books under desk items)
- Different desk chairs (ask coworkers if they want to trade for a day)
- Lighting
    - Really bright daylight bulbs
    - 'Warm light' bulbs
- Different keyboards
- Other ergonomics
- Music
- White/brown noise
- Earplugs
- Establish a policy for interruptions
- Decorations

WORK ROUTINE

- Take breaks to stretch, stand, walk, or meditate
- Try different kinds of work at different times of day
- Create a routine for entering deep focus
- Create a ritual or checklist for ending procrastination and starting work
- Naps
- Snacks
- Co-working
- Collaboration
- Learn keyboard shortcuts for any application you use frequently
- Voice recording (e.g. as notetaking while reading)
- Voice input for computer work (or for your phone)
- Getting Things Done, or a different system for implementing the key principles of attention saving and strategic review
- Time-tracking
- Time-boxing
    - Seriously, spend 15 minutes blocking out hourly plans every day
- Pomodoros (may also help by forcing you to break down tasks)
- LW Study Hall
- Anything on http://lesswrong.com/lw/1sm/akrasia_tactics_review/
- Inbox Zero
- Email filtering
- Process email or other routine tasks in batches
- Virtual assistant
- Remove browser autocomplete suggestions for impulse browsing. It's <shift> <del> with the suggestion highlighted in Chrome.

LEISURE

- Brainstorm a list of endorsed activities to supplant unenjoyable, unrefreshing procrastination
- Walk around, go outside, listen to music, listen to a comedy podcast, meditate, read, do recreational math, sing, dance, exercise, etc.

COMMUTE

- Different routes, weighing stress, safety, scenery, length as you see fit
- Different times of day
- Music/podcasts/audiobooks
- Biking: This can be logistically complicated but still worthwhile. Try seriously thinking for two minutes about what is stopping you from trying it, and whether those obstacles can be removed.

EXERCISE

- Bodyweight workout (various push-ups, sit-ups, pistol squats, etc.)
- Biking
- Running
- Yoga
- A pullup bar or dumbbells
- Dance
- Gym

FOOD

- Write a weekly meal plan
- Find some recipe blogs to follow
- Try a new recipe (bonus points if it's more difficult than usual/from an unfamiliar genre/otherwise stretches your cooking skills)
- Calculate recipe costs
- Go somewhere new or just order something new unusual
- Try snacks/sugar/caffeine at different times of the day
- Reduce/eliminate something (e.g. sugar/caffeine/dairy)
- Try Soylent
- I don't actually know anything about nutrition or dieting; maybe fix that?

MUSIC

- Check whether your favorite musicians (or even ones you only kind of liked before) have released new music
- Find a service where you can listen to music for free with minimal inconvenience (YouTube and Spotify are usable for me, but just barely)

- List artists you've "been meaning to get around to listening to" and use the above to actually do that
- Listen to things outside your usual tastes
- Listen to (internet) radio stations
    - Your local college radio (or any college radio, since they usually stream online) will have a huge variety of programs, some of which should be good
    - Use Google to find a good one; I really like KBAQ for classical
- Try different headphones (go to a store, or ask your friends if you can borrow theirs for a bit)
- Find a "best album of the year" thread from a non-music-related forum; you'll get some pretty diverse picks

OTHER

- Different soap/shampoo/shaving cream/razors/other grooming products
- Different socks
- Barefoot shoes
- Journal
- Gratitude journal
- Comfort zone expansion
- Cold showers
- Meditation
- Various reading, lecture-watching, note-taking, or review strategies
- Alternatives to software and online services you use frequently (whether or not you feel happy with your choice already)
    - http://alternativeto.net and of course http://alternativeto.net/software/alternativeto/
- Watch videos at higher speeds
- HabitRPG, Beeminder, and/or Stickk; brainstorm some goals
    - e.g. pomodoros, Anki reviews/card-making, trying more things, reading, endorsed leisure, meditation, journaling, strategic reviews
- Anki (or spaced repetition more generally)
- The http://tinyhabits.com/ course
- Typing practice to improve speed + accuracy
- An alternative keyboard layout
- Reading practice for speed + comprehension
- Learn to juggle

Habitual Productivity

I was able to maintain high productivity for extended periods of time and achieve some difficult goals. In this and the following posts I will discuss some personality quirks and techniques that helped me do this. This post is fairly self-expository. I claim no originality, this is simply an account of how I operate.

Secret number one: Productivity is a habit of mine. As I mentioned in the previous post, I've been following a similar schedule for years: two days doing social things, five days doing something constructive. Before I turned my efforts towards FAI research, this mainly consistent of programming, writing, and self-education.

This habit was not sufficient to get the high productivity I attained in the last few months, but it was definitely necessary.

I understand that this is not helpful advice: "I'm habitually productive" just passes the buck. "Ah", you ask, "but how did you turn productivity into a habit?" For that, I have an ace up my sleeve:

I *deplore* fun.

Ok, not really. However, I do have a strong aversion to activities that I find unproductive. This aversion is partly innate and partly developed. It first became explicit at the age of nine or ten, when I read *The Phantom Tollbooth*:

> "KILLING TIME!" roared the dog—so furiously that his alarm went off. "It's bad enough wasting time without killing it." And he shuddered at the thought.

- Norton Juster, The Phantom Tollbooth

This quote stuck with me. Time is scarce, and I certainly didn't want to *kill* any.

I developed an explicit distaste for boredom, and went out of my way to avoid it. I kept books near me at all times. I invented stories and thought up new plots when drifting off to sleep. I invented mental puzzles to keep me entertained during class, including a stint in my teens where I worked out the base 12 multiplication tables. Later, I put spare mental cycles towards considering my code, probing edge cases or considering alternative designs (a practice that is no doubt familiar to all programmers).

This distaste broadened as I aged. I grew to realize that I didn't just want to be doing things, I wanted to be doing *useful* things. My disdain started spreading towards other activities, ones that didn't forward my long-term goals. The memories are hazy, and I'm not sure whether this caused or was caused by my naïve resolution to save the world (or a whole tangle of other factors), but I know the two were linked.

Before long, I began to view escapism as a guilty pleasure: fun and addictive, but unsatisfying. Things like hiking and going to parties became almost a chore: I superficially enjoyed them, sure, but I yearned to be elsewhere, doing something *permanent*. Even reading fiction took on a pang of guilt. I valued things that moved me forward, that honed my skills or moved me closer to my terminal goals. I wanted to be *building* things, *improving* things.

This is my first secret weapon: I lost the ability to be satisfied by unproductive activity.

This was not particularly pleasant.

As I got older, I struggled to balance social activities that were supposed to be fun with all of the things that I wanted to learn and build. All forms of entertainment were weighed against their opportunity cost. This wasn't an elegant phase of my life: I was still a teenager, and I yearned for social validation, strong friendships, and adventures just as much as my peers. Trouble was, I was caught in a catch 22: when I squirreled away in my room being "productive" I felt like I was missing out, and when I went outside to have "adventures" I only wanted to be elsewhere. I vacillated wildly for a few years before coming to terms with myself.

These days, I aim to spend about two evenings a week (one on weekdays, one on weekends) doing something that's traditionally fun. I spend the rest of my time doing things that sate my neverending desire to march towards my goals.

It's interesting to note that, in the end, there wasn't really a compromise. The productivity side just flat-out won: I eventually realized that human interaction is necessary for mental health and that a solid social network is invaluable. I don't mean to imply that I engage in social interaction because I've calculated that it's necessary: I *really do* enjoy social interaction, and I *really want* to be able to enjoy it without guilt. Rather, it's more like I've found an excuse that allows me to both enjoy myself and sate the thirst. That said, it's still difficult for me to disengage sometimes.

This is also not the most helpful advice, I realize: I'm good at being productive in part because I'm bad at being satisfied unless my current task forwards my active goals. This isn't exactly something you can practice.

Unless you're into mind hacking, I suppose. (Note: At this point in the post, set your "humor" dials to "dry".)

When I was quite young, one of the guests at our house refused to eat processed food. I remember that I offered her some fritos and she refused. I was fairly astonished, and young enough to be socially inept. I asked, incredulous, how someone could *not like* fritos. To my surprise, she didn't brush me off or feed me banal lines about how different people have different tastes. She gave me the answer of someone who had recently stopped liking fritos through an act of will. Her answer went something like this: "Just start noticing how greasy they are, and how the grease gets all over your fingers and coats the inside of the bag. Notice that you don't want to eat things soaked in that much grease. Become repulsed by it, and then you won't like them either."

Now, I was a stubborn and contrary child, so her ploy failed. But to this day, I still notice the grease. This woman's technique stuck with me. She picked out a *very specific* property of a thing she wanted to stop enjoying and convinced herself that it repulsed her.

If I were *trying* to start hating fun (and I remind you that I'm not trying, because I already do, and that you shouldn't try, because it's no fun) then this is the route I would recommend: Recognize those little discomforts that underlie your escapism, latch on to them, and blow them *completely* out of proportion. (Disclaimer: I am not a mindwizard; I've no doubt there are better ways to change your affections if you're in to mindhacking.)

Note that such mindhacking is a Dark Art which you should not pursue. Side effects may include:

- Experiencing guilt when you should be having a grand old time.
- Attempting to complete hikes as fast as possible so you can get back to what you were working on.
- A propensity to get more tense when you're supposed to be relaxing.
- A tendency to bring books to live concerts so that you can multitask.

Furthermore, I imagine that this can backfire *reaaaly* hard: if you manage to develop a strong revulsion for unproductive activities but *still* can't force yourself to stop browsing reddit (or whatever your vice) then you run a big risk of hitting a willpower-draining death spiral.

So I'm *really* not recommending that you try this mindhack. But if you *already* have spikes of guilt after bouts of escapism, or if you house an arrogant disdain for wasting your time on TV shows, here are a few mantras you can latch on to to help yourself develop a solid hatred of fun (I warn you that these are calibrated for a 14 year old mind and may be somewhat stale):

- When skiing, partying, or generally having a good time, try remembering that this is exactly the type of thing people should have an opportunity to do *after* we stop everyone from dying.
- When doing something transient like watching TV or playing video games, reflect upon how it's not building any skills that are going to make the world a better place, nor really having a lasting impact on the world.
- Notice that if the world is to be saved then it *really does* need to be you who saves it, because everybody else is busy skiing, partying, reading fantasy, or dying in third world countries.

It also helps if you're extraordinarily arrogant and you house a deep-seated belief in civilizational inadequacy.

(You may now disengage your humor shielding.)

I strongly recommend finding a different and preferably healthier route to habitual productivity. The point of this exposition is that *for me*, a quirk of my psychology led me to a schedule where I spend my days doing things that lead towards my goals.

My distaste for other activities is not the thing that is driving me, per se: it has merely pushed me towards a certain lifestyle, it has helped me develop a certain habit. That habit is the foundation for my recent achievements.

If you can structure your life such that productive things are the things that you do *by default*, the things that you do in your free time when you have nothing else on your plate, then you will be in good shape. When "do something that forwards your goals" is the *fallback* plan then it becomes much easier to scale your efforts up.

The way that I built such structure into my own life was pretty personalized and likely unhealthy, but I'm quite content with the end result. So that's my advice for the day: if you can, try to make your default actions useful. Find a way to make productivity habitual.

When forming habits, repetition is very important. If you're trying to be highly productive, consider starting by being a little productive with high regularity. Humans are very habitual creatures, and establishing a habit of completing easier tasks may pay off in the long run.

Even if you start with the easier tasks, though, you're going to need a good chunk of motivation to successfully form a habit of doing things that require effort. In these waters swims Akrasia, a most ancient enemy. I meant to delve more into the sources of my motivation and some tricks I use to avoid akrasia, but I've run out of time. Further posts will follow.

Results from MIRI's December workshop

Last week (Dec. 14-20), MIRI ran its 6th research workshop on logic, probability, and reflection. Writing up mathematical results takes time, and in the past, it's taken quite a while for results from these workshops to become available even in draft form. Because of this, at the December workshop, we tried something new: taking time during and in the days immediately after the workshop to write up results in quick and somewhat dirty form, while they still feel fresh and exciting.

In total, there are seven short writeups. Here's a list, with short descriptions of each. Before you get started on these writeups, you may want to read John Baez's blog post about the workshop, which gives an introduction to the two main themes of the workshop.

Theme 1: Scientific induction in mathematics

One of the main themes was using Bayesian probability to represent uncertainty about mathematical statements. Like human mathematicians, an AI will be able to outright prove or disprove many mathematical statements, but there will also be many that it will be uncertain about, and the obvious thing to try to get a handle on dealing with such uncertainty is to assign a probability to each such statement. This would mean choosing some sort of prior, and then updating on evidence: For example, if you're not sure whether the twin prime conjecture is true, then each time you discover a new twin prime larger than all that you have seen before, you should ever so slightly increase the probability you assign to the conjecture.

But what sort of prior should we choose? That's the problem that people at the December workshop tried to make some progress on. Here is an interesting problem which the best previous proposal, due to Abram Demski, fails on: Suppose that Q(x) is a predicate symbol, and suppose that you find evidence that Q(x) is true of exactly 90% of all numbers between x=1 and x=10^100. Then we would expect that if you plug in some arbitrary number in this range, say n = floor(7^7^7^7 * pi) mod 10^100, then the posterior probability of Q(n), after conditioning on your evidence, would be 0.9. But it turns out that in Demski's proposal, the probability of Q(n) would be approximately 0.5.

To learn more about this, see "**Scientific Induction in Probabilistic Mathematics**", written up by Jeremy Hahn. Jeremy writes: "At the workshop, after pinning down the above example of undesired behaiviour we turned to other proposals for priors. None of the ideas presented are in a polished enough form to be considered a complete proposal, but we are very happy with the progress that was made."

Theme 2: The "procrastination paradox"

The other main theme of the workshop was the Löbian obstacle to self-modifying AI, and more specifically, something we're calling the "procrastination paradox". One approach for constructing a self-modifying AI based on mathematical logic would be to require the AI to prove, whenever it wants to take an action -- be it an external action or making some modification to its own source code -- that this action is "safe". Then

you would naively expect the AI to be able to reason as follows: "The following simple self-modification X is safe, because the modified AI will still only take actions that it has proven to be safe, so it will only take safe actions." But by Löb's theorem, a consistent system T which is as strong as Peano Arithmetic cannot go from "whatever the modified AI does will *have been proven* (in T) to be safe" to "whatever the modified AI does will *in fact* be safe". (See here for discussion of why it seems reasonable to tackle self-modifying AI from the perspective of mathematical logic, even though a realistic AI will certainly need to use probability to reason about the physical world.) In particular, we cannot have a consistent system T as strong as PA that proves "If S is provable in T, then S" for all sentences S.

If we can't have a system T that trusts *itself* in this way, maybe we can have an infinite sequence of consistent systems T_n such that each T_n proves "If S is provable in T_(n+1), then S", for all sentences S? Marcello (Herreshoff) has shown that in fact there is such a sequence, explained in his paper with Eliezer on the Löbian obstacle. However, although it is consistent (meaning that it doesn't prove *everything*; in particular, it doesn't prove 0=1), this system is *unsound* (meaning that it proves *some* things which aren't actually true). We now have an intuition for why this is so, and a corresponding formal proof that any sequence like this must be unsound: Imagine a sequence of agents A_n, using the systems T_n, whose goal it is that a particular button gets pressed -- they don't care *when* it gets pressed, just that it *does* get pressed, eventually. Each of these agents will try to prove that one of its successors presses the button; if it can't find such a proof, then it will press the button itself. Now A_n reasons as follows (in T_n):

> Either my successor, A_(n+1), will press the button, or it won't. If it *doesn't* press the button, then it must have found a proof that the button gets pressed at some later time. This proof is in T_(n+1), and I know that if something is provable in T_(n+1), then it is true. Thus, either way, the button gets pressed -- by my successor, or at some later time.

So A_n has found a proof that the button gets pressed, and therefore doesn't press the button itself. But since this is true for every n, the button never gets pressed -- and therefore, all of the systems T_n are unsound, since they prove a statement which isn't true. Metaphorically, every time the AI thinks about pressing the button, it convinces itself that there is some *future* time when it will stop procrastinating and actually press the button, so *today* it is surely fine to procrastinate -- and in this way, the AI procrastinates forever. For more, including formal proofs, see "**The Procrastination Paradox**", written up by Eliezer.

However, on the positive side, it turns out that if you modify Marcello's construction so that T_n merely proves that T_(n+1) is *consistent*, rather than proving that everything T_(n+1) proves is true, then Marcello's proof not only shows that the T_n are all consistent, but also that they are sound. (This is really easy to see: T_n is PA plus the axiom "psi(n) -> Con(T_(n+1))", for some formula psi(n). Marcello's proof shows that all of these systems T_n are consistent. But then, since T_(n+1) is consistent, "psi(n) -> Con(T_(n+1))" is true, so all axioms of T_n are true and T_n is sound.) Moreover, although there are things that T_n proves that T_(n+1) doesn't -- namely, "Con(T_(n+1))", since if T_(n+1) proved that, it would be inconsistent -- you can construct these T_n so that they all have the same proof-theoretic ordinal, so to the extent that you think that proof-theoretic ordinals are a good measure of the "mathematical strength" of a theory, all of the theories T_n can be said to have the same amount of mathematical strength. Finally, although T_n doesn't prove the Löb schema "If T_(n+1) proves S, then S is true" for *all* sentences S, it does in fact prove it

for a particular class of sentences (the Pi_1 sentences). For details, see "**An infinitely descending sequence of sound theories each proving the next consistent**", written up by Benja Fallenstein, a.k.a. myself.

Is it enough to have the Löb schema for Pi_1 sentences? (These are sentences of the form "forall x. phi(x)", where phi(x) is such that you can write a computer program that takes x as input and outputs whether phi(x) is true or false.) Actually, there's an argument to be made that it might be *exactly right* for tiling agents: It seems to be as much as you can get without running into the procrastination paradox, and it's enough to formalize goals like "The world is not destroyed before time n" (if the world is computable). Moreover, there's a way to combine this with ideas from a different solution to the tiling agents problem, parametric polymorphism (one version of which is described in the tiling agents paper), in order to also handle goals that don't as obviously fit into the Pi_1 mold; we call this hybrid of ideas "**Fallenstein's monster**", written up by Nate Soares a.k.a. So8res, because it looks rather grafted together. (In fact, it was even uglier when we came up with it, because back then we didn't even have the sound version of Marcello's construction yet. I'm hopeful that we will find prettier solutions to the remaining monstrous aspects as well, so that I will soon be able to ask my trusted research assistant, Igor, to dispose of the monster.)

Finally, we usually talk of a set of equations like "T_n = PA + psi(n) -> Con(T_(n+1))" as if they define *one particular* sequence of theories T_n, but it's not obvious that different ways of making this formal lead to logically equivalent theories. In fact, I'd have guessed that in general they don't, but Will Sawin had the opposite intuition and was right; any recursively enumerable set of equations of this type has a unique solution (in the sense that if PA proves that two r.e. lists of theories U_i and V_i are both solutions, then PA proves that they are equivalent). For details, see "**Recursively-defined logical theories are well-defined**", written up by Nisan Stiennon. Interestingly, this turns out to be an equivalent formulation of Löb's theorem!

Odds and ends

In addition to the two main themes, there were some odds and ends: First, it turns out that the 5-and-10 problem, which has been discussed on LW quite a bit, is *almost* a problem for the formalism in the tiling agents paper; the exact system discussed in the paper narrowly avoids the problem, but a simple and natural variant runs into it. For details, see "**The 5-and-10 problem and the tiling agents formalism**", written up by me.

And both last and least, we considered and disposed of a conjecture that a particular version of parametric polymorphism might be able to avoid the problem of losing mathematical strength when an agent using it rewrites itself. This would have been an interesting result if it had been true, so it seems worth recording, but since it turned out to be false, it's not particularly exciting. See "**Decreasing mathematical strength in one formalization of parametric polymorphism**", also written up by me.

Book Review: How Learning Works

As promised, I review and point-by-point summarize **How Learning Works: 7 Research-Based Principles for Smart Teaching** by Susan A. Ambrose, Michael W. Bridges, Michele DiPietro, Marsha C. Lovett, and Marie K. Norman (2010), hereafter *HLW* as I scratch in futility at the sprawling length of this post.

Review

The authors aim to provide "a bridge between research and practice" for teaching and learning, very much in the spirit of Practical Advice Backed by Deep Theories. They concentrate on widely-supported results that are independent of subject matter and environment, so while the discussion is directed towards instructors in K-12 and college classrooms, there are also implications for essentially anyone in a teaching or learning role.

Let me restate that a little more strongly: any student, autodidact or not, would be well-served by internalizing the models and recommendations presented here. Teachers have even less of an excuse not to read the book, which is written very clearly and without sinking to punchy popularization. This is basic stuff, in the best possible way.

Sure, there are more sophisticated ideas out there; there exist subgenres of domain-specific research (especially for math and physics education); you can find diverse perspectives in homeschooling communities or in philosophy of education. There's even some controversy in the depths of the research on some of the points in this book (though for the most part the scope of disagreements is still contained within the boundaries drawn by the authors). But as far as most people need concern themselves, *HLW* is an earnest and accurate if not quite comprehensive account of What We Know about learning.

[I do wish there were a similar account of And How We Think We Know It, looking into common research techniques, metrics of learning outcomes, systematic errors to guard against, reliability of longitudinal studies, statistics about replicability and retractions, and so on, but this isn't it. The book lightly describes methods when it sees fit, and my scattered checks of unfamiliar studies leave me fairly confident that the research does in fact bear the claims the book makes.]

The book organizes research on teaching and learning into seven principles in order to "provide instructors with an understanding of student learning that can help them (a) see why certain teaching approaches are or are not supporting students ’ learning, (b) generate or refine teaching approaches and strategies that more effectively foster student learning in specific contexts, and (c) transfer and apply these principles to new courses." The principles are

1. Students' prior knowledge can help or hinder learning.
2. How students organize knowledge influences how they learn and how they apply what they know.
3. Students' motivation determines, directs, and sustains what they do to learn.
4. To develop mastery, students must acquire component skills, practice integrating them, and know when to apply what they have learned.

5. Goal-directed practice coupled with targeted feedback enhances the quality of students' learning.
6. Students' current level of development interacts with the social, emotional, and intellectual climate of the course to impact learning.
7. To become self-directed learners, students must learn to monitor and adjust their approaches to learning.

Hopefully these ideas are not surprising to you. They are not meant to be; they stand mostly to organize diverse research findings into a coherent model (see principle #2). And if many of those research findings are old news to you as well, I also take that to be a point in favor of the book, and I trust that you will understand why.

Each chapter begins with two stories meant to illustrate the principle, a discussion of the principle itself, a discussion of the research related to that principle, and recommendations that take the principle into account. The chapters are interconnected but stand on their own. If you don't plan to teach, you might get most of your value from Chapters 4, 5, and 7. There's some fluff to the book, but not much. My summary, though long, leaves out the stories and examples, useful repetitions and rephrasings, detailed explanations, and specific recommendations, not to mention descriptions and citations of the relevant studies. I do not consider it a substitute for reading the book, which isn't really that long to begin with.

Before I summarize *HLW*, I'll make a couple brief comparisons. **Why Don't Students Like School: A Cognitive Scientist Answers Questions About How the Mind Works and What It Means for the Classroom** by Daniel T. Willingham (2009) looks pretty similar, down to the format in which chapter titles ask questions which are then answered by Principles of Learning, followed by a discussion of the principle, followed by recommendations for the classroom. It's written at a more popular level, with less discussion of actual research and lots more fluff. Only occasionally does it draw connections directly to a study, rather than use that as the chief mode of exposition (as in *HLW*). Each chapter does have a short annotated bibliography divided into less and more technical texts, which is nice. Willingham comes down strongly in favor of drilling and factual knowledge preceding skill. While that's something I've approvingly polemicized about at some length, it needs a mountain of caveats. In general he optimizes (explicitly, in fact) for counterintuitive punchiness, and it's not always clear how well-supported his advice really is. The organization and coverage feels haphazard to me, but where he hits on topics covered by *HLW*, he seems to agree.

The **25 Principles of Learning** [pdf] from the University of Memphis learning group is a short document with a similar aim: a few sentences describing each principle, a couple sentences describing the implications, and a couple of references. It covers important points that *HLW* addresses only indirectly or that it inexplicably leaves out entirely (spaced repetition, testing, and generation effects, for example). It's worth looking over to fill in those gaps. But it's really "25 Important Findings on Learning": it doesn't give examples, offer very specific advice, or attempt to organize these principles into a causal model of learning. Consider them exercises for the reader.

Summary

1. How Does Students' Prior Knowledge Affect Their Learning?

Students link new ideas and information to what they already know. This can hinder learning in the case of inactive, insufficient, inappropriate, or inaccurate knowledge, but it can also be harnessed to enhance learning.

Research consensus:

- In some ways this is common sense—for example, in the way a mathematics lecture directly relies on definitions and theorems. A student without sufficient background knowledge might still learn to manipulate the symbols, but with more effort, worse retention, worse transfer, and worse ability to explain. But there are also indirect, non-obvious mechanisms at work, in which background knowledge that is not explicitly prerequisite can help learning (as in general knowledge of soccer enhancing recall of arbitrary soccer-match scores).
- *Declarative* knowledge (object-level concepts) and *procedural* knowledge (how and when to apply those concepts) do not always go hand in hand. One without the other is a knowledge gap that can be tricky to spot, especially in self-assessment.
- Existing knowledge needs to be *active* to be effective; activation can be achieved with minor prompts and reminders, as well as questions designed to trigger recall.
- Students may activate existing knowledge that's *inappropriate* (e.g. the colloquial/intuitive meaning of "force" when learning Newtonian physics) or inaccurate. Such activation interferes with learning, leads to incorrect conclusions, and predisposes students to resist conflicting evidence.
- Inaccurate isolated facts can be unlearned through empiricism and explicit refutation. Deeper misconceptions can be extremely persistent, but patient repetition and a long series of small inferential bridges can help.

Strategies for teachers:

- Determine the extent, quality, and nature (e.g. declarative vs. procedural) of students' prior knowledge:
    - Talk to previous instructors
    - Use diagnostic tests
    - Ask self-assessment questions
    - Use brainstorming or concept mapping
    - Look for patterns of error
- Address gaps in prior knowledge:
    - Identify for yourself what knowledge is necessary
    - Remediate insufficient knowledge as determined above
- Activate relevant prior knowledge
    - Explicitly point out connections
    - Use analogies and examples
    - Use exercises that explicitly ask students to use their prior knowledge
- Avoid activating inappropriate prior knowledge:
    - Highlight the boundaries of what knowledge is applicable, either explicitly or with rules of thumb
    - Explicitly identify discipline-specific conventions
    - Show where analogies break down and examples don't generalize
- Help students revise inaccurate knowledge:
    - Ask students to make and test predictions
    - Ask students to justify their reasoning
    - Help students practice using knowledge meant to replace misconceptions
    - Allow sufficient time

2. How Does the Way Students Organize Knowledge Affect Their Learning?

Developing expertise requires rich connections between various facts, concepts, and procedures, organized around abstract principles and causal relationships. Although an expert does not necessarily build such knowledge networks explicitly or consciously, it is possible for a novice learner to deliberately organize knowledge into expert-style structures, improving learning, performance, and retention.

Research:

- The optimal organization of knowledge depends on how that knowledge is to be used. Learning physics in a historical framework has advantages and disadvantages when compared with learning the same physics according to physical principles.
- Students whose knowledge networks (graphs with "pieces of knowledge" as nodes linked by their relationships) are more densely connected will retrieve their knowledge faster and more reliably, and are more likely to notice inconsistencies and contradictions.
- Experts, as a result of their densely connected knowledge networks, process information in coherent *chunks* where novices process individual bits of information (as for chess positions and circuit diagrams). Memorization of digit sequences can be greatly boosted by hierarchical chunking of subsequences. These facts are seen as related.
- Expert knowledge networks have more meaningful connections and deeper organizing principles.
- Students learn better when provided with a structure for organizing information. Causal relationships are especially effective organizing principles.
- Studying worked examples, analogies, and contrasting cases helps students organize their knowledge meaningfully.

Strategies:

- Organize the material:
  - Create a concept map for the material to be taught
  - Identify the knowledge organization best suited to the purpose of learning
- Enhance students' knowledge organization:
  - Explicitly describe the organization of material at each level in the hierarchy of presentation—subject, course, lecture, discussion
  - Use contrasting and boundary cases
  - Explicitly point out deep similarities and other connections
  - Use multiple organizing structures
- Expose students' knowledge organization
  - Ask them to draw a concept map
  - Use a sorting task
  - Look for patterns of mistakes

3. What Factors Motivate Students to Learn?

Students are motivated by the subjective *value* of a goal and by their *expectancy* of success. [You may be reminded of the Procrastination Equation, which also describes

penalties for impulsiveness and delay.] Students may be guided by different goals, and recognizing this can help you foster their motivation.

Research:

- Students who pursue *learning goals*, which emphasize the intrinsic or instrumental value of material, are generally the most motivated and have the best learning outcomes.
- Students may also be guided by *performance goals*, related to their self-image and reputation. These may themselves be performance-approach or performance-avoidant; the former seems to entail a cognitive framework more conducive to learning.
- Work-avoidant goals ("do as little work as possible") can be directly at odds with learning, but are generally context dependent.
- There are, broadly, three broad determinants of subjective value: *attainment* value (satisfaction from mastery or accomplishment), *intrinsic* value, and *instrumental* value. These may mutually reinforce each other.
- To be motivated, a student should expect both their own ability to succeed and for success to bring about a desired outcome.
- Expectancy of success is influenced by the student's past success rate in similar situations, and even more strongly by the reasons the student identifies for their past success or failure. Specifically, attributing success to internal and controllable causes* and failure to controllable but temporary causes increases expectancy. Attributing success to luck and failure to personal inadequacy decreases expectancy. [*Interestingly, the authors make no real distinction here between internal and controllable causes for success, which is a fundamental distinction between the "fixed" vs. "malleable" (which you may know as "growth") mindsets addressed in Chapter 7.]
- Supportive environments also increase motivation.

Strategies:

- Establish value:
    - Connect material to students' interests
    - Provide authentic tasks
    - Show relevance to students' academic lives
    - Show relevance of generalizable skills
    - Identify and reward what you (as the instructor) value
    - Radiate enthusiasm
    - Give students opportunities to reflect on the value of their work
- Build expectancy:
    - Clarify the course goals and your instruction and assessment strategies
    - Identify and set an appropriate level of challenge
    - Help students build success spirals with early challenges
    - Provide feedback on progress
    - Be fair
    - Help students attribute success and failure appropriately
    - Discuss effective study strategies
- Give students flexibility and control in course work to increase both value and expectancy

4. How Do Students Develop Mastery?

Consider a driver changing lanes, making many small motions, visual checks, and mental evaluations fluently and automatically. An expert performs complex tasks with little conscious awareness of the complexity involved. To approach that level of mastery, a novice must not only learn the component skills, but also integrate the skills and know when to apply them.

Research:

- Experts do not necessarily make good teachers: they process information in chunks, they employ shortcuts and skip steps, they perform with automaticity, and they overestimate students' competence. Their unconscious mastery leads to so-called *expert blind spots*.
- Students will perform poorly if their component skills are weak.
- Student performance is greatly improved when instructors identify component skills required for a complex task and target weak ones through practice. A small amount of focused practice on a component skill can have a large impact on performance of the complex task.
- Multitasking degrades performance by way of excess information-processing demands or *cognitive load*. The same applies to combining skills for a complex task, but much more so for novices than for experts.
- Cognitive load can be reduced when learning a complex task by allowing the student to focus on one component skill at a time. It may also be helpful for the instructor to support other aspects of the task while students do their focused practice. This is known as *scaffolding*.
- Another instance of scaffolding effect appears when the instructor presents students with worked examples rather than problems, freeing up cognitive resources to think about principles and techniques.
- Results on drilling component skills in isolation, as compared with practicing the overall task with focus on the components, are mixed. Some skills afford isolated practice better than others. A highly complex but easily divisible task can be learned more effectively by initially practicing the components in isolation, and then progressively combining them.
- Mastery also involves knowing when to apply learned skills outside of the learning context. Doing so is referred to as *transfer*. Transfer occurs rarely and with difficulty, and is worse the more dissimilar the learning and transfer contexts.
- Overspecificity and context-dependence of knowledge hurt transfer; deep understanding of principles and relationships helps transfer. The latter effect can be targeted with structured comparisons and analogical reasoning also help transfer.
- Minor prompts and reminders facilitate transfer, much as they help activate appropriate knowledge (see Chapter 1).

Strategies:

- Expose component skills:
  - Map out your own expert blind spot
  - Enlist help from those with mere conscious competence
  - Talk to others in your discipline
  - Talk to others outside your discipline
  - Explore educational materials
- Reinforce component skills
  - Focus students' attention on the key aspects of the task
  - Diagnose weak or missing component skills

    - Provide isolated practice of those skills.
- Build fluency and facilitate integration of skills
    - Give students practice exercises explicitly to increase automaticity
    - Temporarily constrain the scope of the task
    - Explicitly include integration in performance criteria
- Facilitate transfer:
    - Discuss conditions of applicability
    - Give exercises explicitly about conditions of applicability
    - Provide opportunities to practice in diverse contexts
    - Use hypothetical scenarios for practice questions
    - Ask students to generalize to abstract principles
    - Identify deep features using comparisons
    - Prompt students to retrieve relevant knowledge

5. What Kinds of Practice and Feedback Enhance Learning?

Practice is often misguided and feedback poorly timed, insufficient, or unfocused. To be effective, practice should be directed by goals and coupled with targeted feedback.

Research:

- Learning can be predicted by time in *deliberate practice*, which is marked by being directed toward a specific goal and an appropriate level of challenge. [I've often heard deliberate practice described with an emphasis on mindful attention, in contrast with practice in a flow state (for example in an article by Ericsson himself—the last paragraph before "Future Directions"), but the authors questionably suggest that flow is a sign of appropriate challenge. For motivation, perhaps it is, but I would argue not so for deliberate practice.]
- Clearly specified performance criteria can help direct students' practice.
- Learning is hampered by either insufficient or excessive challenge.
- The success of individual tutoring is largely driven by the ability to tailor challenges to a level appropriate to deliberate practice.
- An instructor can improve learning outcomes with difficult tasks by adding structure and support to bring it within the bounds of the student's competence. This can consist of guidance by the instructor, or of written prompts and checklists. (C.f. "scaffolding" in Chapter 4.)
- The benefits of deliberate practice accrue gradually with increasing time spent practicing; both students and teachers underestimate the time needed.
- The effectiveness of feedback is determined by both content and timing. It should communicate progress and direct subsequent effort, and it should be supplied when students can best use it.
- Feedback that identifies specific items that need improvement will aid learning more than will a mere indication of error.
- Unfocused feedback can be counterproductive by overwhelming the student and failing to direct effort well.
- Generally, more frequent and more rapid feedback is better for learning. Delayed feedback can be useful in helping students learn to recognize and correct their own errors.

Strategies:

- Establish goals:
  - Be explicit about course goals, and phrase them in terms of capabilities rather than knowledge
  - Use a rubric to communicate performance criteria
  - Give contrasting examples of high and low quality work
  - Progressively refine goals
- Encourage deliberate practice:
  - Assess prior knowledge to set an appropriate challenge
  - Create many chances to practice
  - Build scaffolding into assignments
  - Set expectations about practice
- Target feedback:
  - Look for patterns of errors
  - Use prioritized feedback to direct student efforts
  - Give feedback on strengths and weaknesses
  - Allow frequent opportunities for feedback
  - Provide feedback at the group level, potentially in real-time
  - Require peer feedback on assignments
  - Require students to describe how they incorporated feedback

6. Why Do Student Development and Course Climate Matter for Student Learning?

People vary not just intellectually, but also socially and emotionally. Students' identities may be entangled with the course material and environment in complicated ways that often go unrecognized. A student's entire state—not just the intellect—interacts with the social, emotional, and intellectual climate of the course to impact learning, for better or for worse. [When I saw this chapter title, I had a vague worry that it would seem out of place, a perfunctory nod to diversity studies or something. I'm still not entirely comfortable with parts of the treatment here, but the above premise is sound.]

Research:

- The research involved in this first section is of a different nature from the rest of the text. In the first part, the authors seek to describe student development, and cite a model which characterizes developmental changes into seven dimensions: developing competence, managing emotions, developing autonomy, establishing identity, freeing interpersonal relationships, developing purpose, and developing identity. They then cite research characterizing intellectual developments in terms of stages: duality, multiplicity, relativism, and commitment. Similarly, stages for social development. The point is that people can have a lot of different implicit and explicit beliefs, modes of communication, and ways of processing new information, which they can't just switch off and homogenize when they enter a classroom, and that people have done a lot of work to attempt to enumerate and connect these things. [I think the discussion here is the weakest part of the book, and I'd be interested in better resources on the subject, if they exist.]
- For course climate, they describe a classification in terms of whether an environment is *marginalizing* or *centralizing* (describing how the perspectives of groups might be discouraged or welcomed), and whether this occurs *implicitly* or

explicitly. Implicitly marginalizing classrooms are the most common of the four quadrants.

- In implicitly marginalizing environments (i.e. without overt exclusion or hostility towards outgroups), individuals may suffer an accumulation of *micro-inequities* that over time has a large impact on learning. A number of studies have found that perceptions of a marginalizing climate are negatively correlated with learning and career outcomes. The authors identify four important channels for marginalization: stereotypes, tone, faculty-student interactions, and content.
- The activation (in the sense of Chapter 1) of stereotypes can influence learning, generally impairing performance; this effect is known as *stereotype threat*. The activation does not have to be a result of explicitly invoking the stereotype; implicit communication of assumptions or apparently innocuous comments also have effects.
- The immediate mechanism for stereotype threat seems to be a disruptive emotional reaction; this as opposed to self-efficacy or self-esteem being depressed or otherwise brought in line with the stereotype. The effect does not require any belief in the stereotype. There are deeper nuances as well as strategies for mitigating the effect in the literature.
- Perceived hostility or expectations of failure in stereotypes can decrease motivation and drive students from a discipline.
- A positive, constructive, and encouraging tone in discussions and syllabi improves student motivation and behavior. (This in contrast to punitive, critical, or demeaning tone.)
- Perceived positive faculty attitudes towards and interactions with undergrads are correlated with higher rates of graduate education and better self-reported learning outcomes. Faculty availability is a major factor in students' academic decisions.
- Course content itself in its orientation towards inclusiveness can have cognitive, motivational, and socio-emotional effects on learning.

Strategies:

- Promote intellectual development:
    - Make uncertainty, ambiguity, and complexity safe
    - Resist a single right answer
    - Incorporate use of evidence into performance criteria
- Promote social development:
    - Examine your assumptions about your students
    - Be mindful of accidental cues regarding stereotypes
    - Do not ask individuals to speak for an entire group
    - Recognize students as individuals.
- Promote an inclusive climate:
    - Be a model for inclusive language, behavior, and attitudes
    - Use multiple and diverse examples
    - Establish and reinforce ground rules for interaction
    - Make sure course content does not marginalize students
    - Use the syllabus and first day of class to establish climate
    - Set up processes to get feedback on the climate
    - Anticipate and prepare for sensitive issues
    - Address tensions early
    - Turn discord and tension into a learning opportunity
    - Facilitate and model active listening.

7. How Do Students Become Self-Directed Learners?

As one progresses in academic and professional life, one takes progressively more responsibility for one's own learning. The jump between high school and college can be especially jarring in this regard. *Metacognition*, "the process of reflecting on and directing one's own thinking," becomes increasingly important, but falls outside the scope of most instruction. Still, to effectively direct their own learning, students must learn and practice an array of metacognitive skills.

Research:

- One model represents metacognition as a continuously looping cycle of task assessment, evaluation of strengths and weaknesses, planning, execution and simultaneous monitoring, and reflection; all of these five steps are informed by a student's beliefs about intelligence and learning.
- Assessing the task is not always natural or obvious to students (essay prompts are often ignored; learning goals are not always clear).
- People are poor judges of their own knowledge and skills, tending to overestimate their abilities more the weaker they are.
- Novices spend little time in the planning phase of the cycle relative to experts in physics, math, and writing. Novice plans are often poorly matched to the task.
- Students who naturally and continuously monitor their performance and understanding learn better.
- Students can be taught to self-monitor, and this also improves learning.
- Monitoring alone is not sufficient; novice problem solvers will continue to use a strategy after it has failed (and certainly after it has proven modestly successful and familiar but not optimal).
- Students who believe their intelligence is malleable rather than fixed are more likely to learn and perform well.
- Moreover, the "malleable" perspective can be promoted by external influences, still leading to better performance.

Strategies:

- Promote task assessment:
    - Be more explicit about assignments than you think is necessary
    - Tell students what you do not want
    - Check students' understanding of the task in their own words
    - Provide a rubric
- Promote self-evaluation:
    - Give timely feedback
    - Provide opportunities for self-assessment.
- Promote planning:
    - Have students implement a plan you provide
    - Have students implement their own plan
    - Make planning the central goal of the assignment.
- Promote self-monitoring:
    - Provide simple heuristic questions for self-evaluation
    - Have students do guided self-assessments
    - Require students to reflect on and annotate their own work
    - Use peer review

- Promote reflection and adjustment:
    - Prompt students to reflect on their performance
    - Prompt students to analyze effectiveness of study skills
    - Present multiple strategies
    - Create assignments that focus on strategizing
- Promote useful beliefs about intelligence and learning:
    - Address these beliefs directly
    - Broaden students' understanding of learning
    - Help students set realistic expectations
- Promote metacognition:
    - Model your metacognitive process for your students
    - Scaffold students in their metacognitive processes

Conclusion: Applying the Seven Principles to Ourselves

The authors turn their principles inward and discuss learning to teach. For the most part this is a restatement of the principles with no particularly new insights in their application to teaching, but there are interesting comments regarding the first few:

- Many teachers were formerly atypically successful students, and their prior knowledge can lead to distorted expectations; accordingly, many of the recommendations involve gathering data about the students.
- The organization of this book into principles is itself a deliberate application of the second principle.
- For motivation, the authors try to connect the content of the course with what every teacher really cares about: efficiency. They also suggest focusing on one or two aspects of teaching in a given semester, in order to build up small successes in improving teaching.
- In terms of mastery, practice and feedback, climate, and metacognition, teaching is not so different from any other skill.

Appendices

HLW has eight appendices on tools mentioned throughout the book, with a reiteration of their nature and utility, and most importantly, example checklists and worksheets. These are

- Student self-assessment
- Concept maps
- Rubrics
- Learning objectives
- Ground rules [for discussion]
- Exam wrappers [for promoting metacognition on graded exams]
- Checklists
- Reader response/peer review

These alone would have been an improvement over most teaching materials I grew up with.

Fascists and Rakes

Cross-posted from my blog

It feels like most people have a moral intuition along the lines of "you should let people do what they want, unless they're hurting other people". We follow this guideline, and we expect other people to follow it. I'll call this the permissiveness principle, that behaviour should be permitted by default. When someone violates the permissiveness principle, we might call them a fascist, someone who exercises control for the sake of control.

And there's another moral intuition, the harm-minimising principle: "you should not hurt other people unless you have a good reason". When someone violates harm-minimisation, we might call them a rake, someone who acts purely for their own pleasure without regard for others.

But sometimes people disagree about what counts as "hurting other people". Maybe one group of people believes that tic-tacs are sentient, and that eating them constitutes harm; and another group believes that tic-tacs are not sentient, so eating them does not hurt anyone.

What should happen here is that people try to work out exactly what it is they disagree about and why. What actually happens is that people appeal to permissiveness.

Of course, by the permissiveness principle, people should be allowed to believe what they want, because holding a belief is harmless as long as you don't act on it. So we say something like "I have no problem with people being morally opposed to eating tic-tacs, but they shouldn't impose their beliefs on the rest of us."

Except that by the harm-minimising principle, those people probably *should* impose their beliefs on the rest of us. Forbidding you to eat tic-tacs doesn't hurt you much, and it saves the tic-tacs a lot of grief.

It's not that they disagree with the permissiveness principle, they just think it doesn't apply. So appealing to the permissiveness principle isn't going to help much.

I think the problem (or at least part of it) is, depending how you look at it, either double standards or not-double-enough standards.

I apply the permissiveness principle "unless they're hurting other people", which really means "unless I think they're hurting other people". I want you to apply the permissiveness principle "unless they're hurting other people", which *still* means "unless I think they're hurting other people".

Meanwhile, you apply the permissiveness principle unless *you* think someone is hurting other people; and you want me to apply it unless *you* think they're hurting other people.

So when we disagree about whether or not something is hurting other people, I think you're a fascist because you're failing to apply the permissiveness principle; and you think I'm a rake because I'm failing to apply the harm-minimisation principle; or vice-versa. Neither of these things is true, of course.

It gets worse, because once I've decided that you're a fascist, I think the *reason we're arguing* is that you're a fascist. If you would only stop being a fascist, we could get along fine. You can go on thinking tic-tacs are sentient, you just need to stop being a fascist.

But you're not a fascist. The *real* reason we're arguing is that you think tic-tacs are sentient. You're acting exactly as you should do if tic-tacs were sentient, but they're not. I need to stop treating you like a fascist, and start trying to convince you that tic-tacs are not sentient.

And, symmetrically, you've decided I'm a rake, which isn't true, and you've decided that that's why we're arguing, which isn't true; we're arguing because I think tic-tacs aren't sentient. You need to stop treating me like a rake, and start trying to convince me that tic-tacs are sentient.

I don't expect either of us to actually convince the other, very often. If it was that easy, someone would probably have already done it. But at least I'd like us both to acknowledge that our opponent is neither a fascist nor a rake, they just believe something that isn't true.

Even Odds

(Cross-posted on my personal blog, which has LaTeX, and is easier to read.)

Let's say that you are are at your local less wrong meet up and someone makes some strong claim and seems very sure of himself, "blah blah blah resurrected blah blah alicorn princess blah blah 99 percent sure." You think he is probably correct, you estimate a 67 percent chance, but you think he is way over confident. "Wanna bet?" You ask.

"Sure," he responds, and you both check your wallets and have 25 dollars each. "Okay," he says, "now you pick some betting odds, and I'll choose which side I want to pick."

"That's crazy," you say, "I am going to pick the odds so that I cannot be taken advantage of, which means that I will be indifferent between which of the two options you pick, which means that I will expect to gain 0 dollars from this transaction. I wont take it. It is not fair!"

"Okay," he says, annoyed with you. "We will both write down the probability we think that I am correct, average them together, and that will determine the betting odds. We'll bet as much as we can of our 25 dollars with those odds."

"What do you mean by 'average' I can think of at least four possibilities. Also, since I know your probability is high, I will just choose a high probability that is still less than it to maximize the odds in my favor regardless of my actual belief. Your proposition is not strategy proof."

"Fine, what do you suggest?"

You take out some paper, solve some differential equations, and explain how the bet should go.

Satisfied with your math, you share your probability, he puts 13.28 on the table, and you put 2.72 on the table.

"Now what?" He asks.

A third meet up member takes quickly takes the 16 dollars from the table and answers, "You wait."

I will now derive a general algorithm for determining a bet from two probabilities and a maximum amount of money that people are willing to bet. This algorithm is both strategy proof and fair. The solution turns out to be simple, so if you like, you can skip to the last paragraph, and use it next time you want to make a friendly bet. If you want to try to derive the solution on your own, you might want to stop reading now.

First, we have to be clear about what we mean by strategy proof and fair. "Strategy proof" is clear. Our algorithm should ensure that neither person believes that they can increase their expected profit by lying about their probabilities. "Fair" will be a little harder to define. There is more than one way to define "fair" in this context, but there is one way which I think is probably the best. When the players make the bet, they both will expect to make some profit. They will not both be correct, but they will both

believe they are expected to make profit. I claim the bet is fair if both players expect to make the same profit on average.

Now, lets formalize the problem:

Alice believes S is true with probability p. Bob believes S is false with probability q. Both players are willing to bet up to d dollars. Without loss of generality, assume p+q>1. Our betting algorithm will output a dollar amount, f(p,q), for Alice to put on the table and a dollar amount, g(p,q) for Bob to put on the table. Then if S is true, Alice gets all the money, and if S is false, Bob gets all the money.

From Alice's point of view, her expected profit for Alice will be p(g(p,q))+(1-p)(-f(p,q)).

From Bob's point of view, his expected profit for Bob will be q(f(p,q))+(1-q)(-g(p,q)).

Setting these two values equal, and simplifying, we get that (1+p-q)g(p,q)=(1+q-p)f(p,q), which is the condition that the betting algorithm is fair.

For convenience of notation, we will define h(p,q) by h(p,q)=g(p,q)/(1+q-p)=f(p,q)/(1+p-q).

Now, we want to look at what will happen if Alice lies about her probability. If instead of saying p, Alice were to say that her probability was r, then her expected profit would be p(g(r,q))+(1-p)(-f(r,q)), which equals p(1+q-r)h(r,q)+(1-p)(-(1+r-q)h(r,q))= (2p-1-r+q)h(r,q).

We want this value as a function of r to be maximized when r=p, which means that -h+(2r-1-r+q)(dh/dr)=0.

Separation of variables gives us (1/h)dh=1/(-1+r+q)dr,

which integrates to ln(h)=C+ln(-1+r+q) at r=p,

which simplifies to h=e^C(-1+r+q)=e^C(-1+p+q).

This gives the solution f(p,q)=e^C(-1+p+q)(1+p-q)=e^C(p^2-(1-q)^2) and g(p,q)=e^C(-1+p+q)(1+q-p)=e^C(q^2-(1-p)^2).

It is quick to verify that this solution is actually fair, and both players' expected profit is maximized by honest reporting of beliefs.

The value of the constant multiplied out in front can be anything, and the most either player could ever have to put on the table is equal to this constant. Therefore, if both players are willing to bet up to d dollars, we should define e^C=d.

Alice and Bob are willing to bet up to d dollars, Alice thinks S is true with probability p, and Bob thinks S is false with probability q. Assuming p+q>1, Alice should put in d(p^2-(1-q)^2), while Bob should put in d(q^2-(1-p)^2). I suggest you use this algorithm next time you want to have a friendly wager (with a rational person), and I suggest you set d to 25 dollars and require both players to say an odd integer percent to ensure a whole number of cents.

Humans can drive cars

There's been a lot of fuss lately about Google's gadgets. Computers can drive cars - pretty amazing, eh? I guess. But what amazed me as a child was that *people* can drive cars. I'd sit in the back seat while an adult controlled a machine taking us at insane speeds through a cluttered, seemingly quite unsafe environment. I distinctly remember thinking that something about this just doesn't add up.

It looked to me like there was just no adequate mechanism to keep the car on the road. At the speeds cars travel, a tiny deviation from the correct course would take us flying off the road in just a couple of seconds. Yet the adults seemed pretty nonchalant about it - the adult in the driver's seat could have relaxed conversations with other people in the car. But I knew that people were pretty clumsy. I was an ungainly kid but I knew even the adults would bump into stuff, drop things and generally fumble from time to time. Why didn't that seem to happen in the car? I felt I was missing something. Maybe there were magnets in the road?

Now that I am a driving adult I could more or less explain this to a 12-year-old me:

1. Yes, the course needs to be controlled very exactly and you need to make constant tiny course corrections or you're off to a serious accident in no time.

2. Fortunately, the steering wheel is a really good instrument for making small course corrections. The design is somewhat clumsiness-resistant.

3. Nevertheless, you really are just one misstep away from death and you need to focus intently. You can't take your eyes off the road for even one second. Under good circumstances, you can have light conversations while driving but a big part of your mind is still tied up by the task.

4. People can drive cars - but only just barely. You can't do it safely even while only mildly inebriated. That's not just an arbitrary law - the hit to your reflexes substantially increases the risks. You can do pretty much all other normal tasks after a couple of drinks, but not this.

So my 12-year-old self was not completely mistaken but still ultimately wrong. There are no magnets in the road. The explanation for why driving works out is mostly that people are just somewhat more capable than I'd thought. In my more sunny moments I hope that I'm making similar errors when thinking about artificial intelligence. Maybe creating a safe AGI isn't as impossible as it looks to me. Maybe it isn't beyond human capabilities. Maybe.

Edit: I intended no real analogy between AGI design and driving or car design - just the general observation that people are sometimes more competent than I expect. I find it interesting that multiple commenters note that they have also been puzzled by the relative safety of traffic. I'm not sure what lesson to draw.

[link] Why Self-Control Seems (but may not be) Limited

Another attack on the resource-based model of willpower, Michael Inzlicht, Brandon J. Schmeichel and C. Neil Macrae have a paper called "Why Self-Control Seems (but may not be) Limited" in press in *Trends in Cognitive Sciences*. Ungated version here.

Some of the most interesting points:

- Over 100 studies appear to be consistent with self-control being a limited resource, but generally these studies do not observe resource depletion directly, but infer it from whether or not people's performance declines in a second self-control task.
- The only attempts to directly measure the loss or gain of a resource have been studies measuring blood glucose, but these studies have serious limitations, the most important being an inability to replicate evidence of mental effort actually affecting the level of glucose in the blood.
- Self-control also seems to replenish by things such as "watching a favorite television program, affirming some core value, or even praying", which would seem to conflict with the hypothesis inherent resource limitations. The resource-based model also seems evolutionarily implausible.

The authors offer their own theory of self-control. One-sentence summary (my formulation, not from the paper): "Our brains don't want to only work, because by doing some play on the side, we may come to discover things that will allow us to do even more valuable work."

- Ultimately, self-control limitations are proposed to be an exploration-exploitation tradeoff, "regulating the extent to which the control system favors task engagement (exploitation) versus task disengagement and sampling of other opportunities (exploration)".
- Research suggests that cognitive effort is inherently aversive, and that after humans have worked on some task for a while, "ever more resources are needed to counteract the aversiveness of work, or else people will gravitate toward inherently rewarding leisure instead". According to the model proposed by the authors, this allows the organism to both focus on activities that will provide it with rewards (exploitation), but also to disengage from them and seek activities which may be even more rewarding (exploration). Feelings such as boredom function to stop the organism from getting too fixated on individual tasks, and allow us to spend some time on tasks which might turn out to be even more valuable.

The explanation of the actual proposed psychological mechanism is good enough that it deserves to be quoted in full:

> Based on the tradeoffs identified above, we propose that initial acts of control lead to shifts in motivation away from “have-to” or “ought-to” goals and toward “want-to” goals (see Figure 2). “Have-to” tasks are carried out through a sense of duty or contractual obligation, while “want-to” tasks are carried out because they are personally enjoyable and meaningful [41]; as such, “want-to” tasks feel easy to perform and to maintain in focal attention [41]. The distinction between “have-

to” and “want-to,” however, is not always clear cut, with some “want-to” goals (e.g., wanting to lose weight) being more introjected and feeling more like “have-to” goals because they are adopted out of a sense of duty, societal conformity, or guilt instead of anticipated pleasure [53].

According to decades of research on self-determination theory [54], the quality of motivation that people apply to a situation ranges from extrinsic motivation, whereby behavior is performed because of external demand or reward, to intrinsic motivation, whereby behavior is performed because it is inherently enjoyable and rewarding. Thus, when we suggest that depletion leads to a shift from “have-to” to “want-to” goals, we are suggesting that prior acts of cognitive effort lead people to prefer activities that they deem enjoyable or gratifying over activities that they feel they ought to do because it corresponds to some external pressure or introjected goal. For example, after initial cognitive exertion, restrained eaters prefer to indulge their sweet tooth rather than adhere to their strict views of what is appropriate to eat [55]. Crucially, this shift from “have-to” to “want-to” can be offset when people become (internally or externally) motivated to perform a “have-to” task [49]. Thus, it is not that people cannot control themselves on some externally mandated task (e.g., name colors, do not read words); it is that they do not feel like controlling themselves, preferring to indulge instead in more inherently enjoyable and easier pursuits (e.g., read words). Like fatigue, the effect is driven by reluctance and not incapability [41] (see Box 2).

Research is consistent with this motivational viewpoint. Although working hard at Time 1 tends to lead to less control on “have-to” tasks at Time 2, this effect is attenuated when participants are motivated to perform the Time 2 task [32], personally invested in the Time 2 task [56], or when they enjoy the Time 1 task [57]. Similarly, although performance tends to falter after continuously performing a task for a long period, it returns to baseline when participants are rewarded for their efforts [58]; and remains stable for participants who have some control over and are thus engaged with the task [59]. Motivation, in short, moderates depletion [60]. We suggest that changes in task motivation also mediate depletion [61].

Depletion, however, is not simply less motivation overall. Rather, it is produced by lower motivation to engage in “have-to” tasks, yet higher motivation to engage in “want-to” tasks. Depletion stokes desire [62]. Thus, working hard at Time 1 increases approach motivation, as indexed by self-reported states, impulsive responding, and sensitivity to inherently-rewarding, appetitive stimuli [63]. This shift in motivational priorities from “have-to” to “want-to” means that depletion can increase the reward value of inherently-rewarding stimuli. For example, when depleted dieters see food cues, they show more activity in the orbitofrontal cortex, a brain area associated with coding reward value, compared to non-depleted dieters [64].

See also: Kurzban et al. on opportunity cost models of mental fatigue and resource-based models of willpower; Deregulating Distraction, Moving Towards the Goal, and Level Hopping.

Soft Paternalism in Parenting

Reading the recently featured Beware of Trivial Inconveniences I realized that this is the method that makes Say Yes really work and thus this is Practical Advice Backed By Deep Theories.

The trick of saying "yes" instead of "no" is *not* to say less often "no" at the cost at allowing things when you say "yes". That just trades the stress of saying "no" (staying consequent despite a clash of wills) against the effort to fulfill, monitor, pay or clean up after the "yes".

Soft paternalism applied to parenting means saying "Yes, but" or "Yes, later" or "Yes, if". This signals to the child that you understand his/her wish but also supplies some context the child may not be aware of. It reduces your cost of saying "yes" at the expense of a cost to cash in the "yes" for the child.

Disclaimer: This 'cost reversal' works if

- the condition is no artificial construction to make the "yes" into an effective "no" (in which case the child will learn this pattern of disguised "no" and might e.g. feel cheated. Though this may still be more polite than saying plain "no".
- The condition/context for the "yes" provides real information for the child.
- The child is old enough to at least grasp the concept of a condition (is in its Zone of Proximal Development)

Examples:

I use this pattern...

For my oldest (10) when he wants to do some larger/elaborate projects and e.g. asks "may I organize event X" I don't want to stiffle his motivation to show responsibility, learn required tasks and socialize. But I also don't want to do significant parts of this. So I e.g. say: "Yes, but you have to consult the calendar for a time, write the invitation yourself and clean up afterwards".

For my second oldest (7) if he wants some book or other piece of parent stuff like bowls I request: "yes, but put it back afterwards".

This will not work on his younger brother (5) who is not yet disciplined enough to remember to put things back afterwards. For him a limitation like "yes, but not now; after I have done the dishes" is more applicable. It a) places a condition he can monitor (or is motivated to monitor) and b) it moves the fulfillment into a position where I am more willing to do/allow/give it and c) when he forgets about it I got rid of it at no cost.

When applying this to the youngest (2) the conditions need to be most simple as he doesn't understand the condition yet. He hears the "yes but bla bla" and recognizes that it somehow means that he doesn't get it (yet). His reaction is mostly the same as to a "no" but I can talk some more to him to make clear that it is no "no" (and his brothers support this: "don't cry, you will get it"). So this montivates him to learn the linguistic concept of a conditional.

One still has to be consequent in standing behind the context/condition and not be turned around to a plain "yes".

And for older children you still have to be careful what you say. This is already hard for the oldest who is constantly trying to optimize for the conditions or bend the wording of the conditions.

Such a semi-permissive parenting is also called Authoritative Parenting and generally more successful in preparing children for adult life. See e.g.

http://www.parentingscience.com/authoritative-parenting-style.html (The authoritative parenting style: Warmth, rationality, and high standards: A guide for the science-minded parent)

http://www.philly.com/philly/blogs/healthy_kids/You-say-yesI-say-no-how-parenting-style-may-affect-teens-behaviors-.html

EDIT: Fixed some typos of this obviously still occassionally read post.

Decision Auctions aka "How to fairly assign chores, or decide who gets the last cookie"

After moving in with my new roomies (Danny and Bethany of Beeminder), I discovered they have a fair and useful way of auctioning off joint decisions. It helps you figure out how much you value certain chores or activities, and it guarantees that these decisions are worked out in a fair way. They call it "yootling", and wrote more about it here.

A quick example (Note: this only works if all participants are of the types of people who consider this sort of thing a Good Idea, and not A Grotesque Parody of Caring or whatnot):

Use Case: Who Picks up the Kids from Grandma's?

D and B are both busy working, but it's time to pick up the kids from their grandparents house. They decide to yootle for it.

B bids $100 (In a regular Normal Person exchange, this would be like saying "I'm elbows deep in code right now, and don't want to break flow. I'd really rather continue working right now, but of course I'll go if it's needed.")

D bids $15 (In a regular Normal Person exchange this would be like saying "I don't mind too much, though I do have other things to do now...")

So D "wins" the bid, and B pays him $15 to go get the kids from their grandma's.

Of course.... it would be a pain in the butt to constantly be paying each other, so instead they have a 10% chance of paying 10x the amount, and a 90% chance to pay nothing, using a random number generator.

This is made easier by the fact that we have a bot to run this, but before that they would use the high-tech solution of Holding Up Fingers.

> We may do this multiple times per day, whenever there’s a good that we have shared ownership of and one of us wants to offload their shares onto the other person. The goods can be anything, e.g. the last brownie, but they’re more often “bads” like who will get up in the middle of the night with a vomiting child, or who will book plane tickets for a trip.
>
> We find this an elegant means of assigning loathed tasks. The person who minded least winds up doing the chore, but gets compensated for it at a price that by their own estimation was fair.

Some other ways it can be implemented:

Joint purchase auction

The decision auction and variants are about allocating shared or partially shared resources to one person or the other, or picking one person to do something. Once in a while you have the opposite problem: deciding on a joint purchase.

Suppose Danny thinks we need a new sofa (this is very hypothetical). I think the one we have is just fine thank you. After some discussion I concede that it would be nice to have a sofa that was less doggy. Danny, being terribly excited about getting a new sofa does a bunch of research and finds his ideal sofa. I think it is a bit overpriced considering it is going to be a piece of gymnastics equipment for the kids for the next 6 years. Conflict ensues! I could bluff that I'm not interested in a new sofa at all and that he can buy it himself if he wants it that badly. But he probably doesn't want it that bad, and I do want it a little. If only we could buy the sofa conditional on our combined utility for it exceeding the cost, and pay in proportion to our utilities to boot. Well, thanks to separate finances and the magic of mechanism design, we can! We submit sealed bids for the sofa and buy it if the sum of our bids is enough. (And, importantly, commit to not buying it for at least a year otherwise.) Any surplus is redistributed in proportion to our bids. For example, if Danny bid $80 and I bid $40 to buy a hundred dollar sofa, then we'd buy it, with Danny chipping in twice as much as me, namely $67 to my $33.

Generosity without sacrificing social efficiency

"The payments are simply what keep us honest in assessing that."
If you're thinking "how mercenary all this is!" then, well, I'm unclear how you made it this far into this post. But it's not nearly as cold as it may sound. We do nice things for each other all the time, and frequently use yootling to make sure it's socially efficient to do so. Suppose I invite Danny to a sing-along showing of Once More With Feeling (this may or may not be hypothetical) and Danny doesn't exactly want to go but can see that I have value for his company. He might (quite non-hypothetically) say "I'll half-accompany you!" by which he means that he'll yootle me for whether he goes or not. In other words, he magnanimously decides to treat his joining me as a 50/50 joint decision. If I have greater value for him coming than he has for not coming, then I'll pay him to come. But if it's the other way around, he will pay me to let him off the hook. We don't actually care much about the payments, though those are necessary for the auction to work. We care about making sure that he comes to the Buffy sing-along if and only if my value for his company exceeds his value for staying home. The payments are simply what keep us honest in assessing that. The increased fairness — the winner sharing their utility with the loser — is icing.

European Community Weekend in Berlin

The Berlin Meetup Group is organizing the first European community meetup. We are planning a fun weekend with a focus on bringing the LessWrong community closer together. As a treat, some participants offer rationality exercises and workshops.

If you like your local meetup we hope you will like this too. It is similar, but bigger: You will get to meet and exchange ideas with a diverse set of awesome people from all across Europe. And if you don't have a meetup nearby or didn't get around to participating yet, this is a great opportunity to get in touch with the rest of the community.

The community weekend will take place April 11-13, from Friday evening to Sunday early afternoon, in the Odyssee Hostel in Berlin. The cost is 70 € including accommodation and breakfast. A conference room with a projector and wifi will also be available during daytime.

The event is participant driven so you (yes, you) are very welcome to help with the content. Please contact Tristan (wegnertristan - at - googlemail - dot - com) for planning and scheduling.

Friday evening we will welcome everyone around 17:00 and proceed to have dinner together. Sleeping quarters (shared rooms) will already be available from 16:00. The evening can be used for getting to know each other better or exploring the city in small groups. We also made sure that there are plenty of cultural, party and sightseeing offers available throughout the city.

Saturday after breakfast, there will be a semi-structured program in the conference room. Offers will include a workshop for habit implementation, a science based introduction to meditation along with a practice session, and clicker training. We also encourage everyone to take part in lightning talks, where you are invited to give a short 5min presentation on any topic you are passionate about.

Saturday evening can be further used to nerd out, explore the nightlife or participate in an exercise in social comfort zone expansion.

Sunday will have room for additional projects. We will say goodbye in the early afternoon to give us enough time to head home. Of course everyone is free to prolong the stay in the greatest capital of Germany.

By the end of the event we hope that we will have made new friends, exchanged ideas, received valuable feedback and started a joint world optimization project or two.

To sign up, email John (johncryptfrink - at - gmail - dot - com), preferably before the first of February. The number of participants is limited, so register early to make sure you can participate.

Looking forward to seeing you

John, Tristan, Alexander, Matthias, Christian… & everyone else from the Berlin LessWrong meetup

UPDATE: We're now at slightly over 40 participants and can't accept any more people. Please mail John anyway if you'd have wanted to come! Things may still work out if someone else can't make it after all - and it will help us have a better idea of how many people to plan for next time.

Things I Wish They'd Taught Me When I Was Younger: Why Money Is Awesome

There are some things money can't buy. They are the exceptions that prove the rule.

For the pedants, to say something is an exception that proves the rule is to say that when you look at the exceptions, they're so unusual that it reinforces the point that the rule is generally valid even though it isn't universally valid. In the case of money, there's a reason people don't say things like "there are some things hand-knit scarves can't be bartered for" or "Hand-knit scarves can't be bartered for happiness."

Eliezer once described the sequences as the letter he wishes he could have written to his former self. When I think of the letter I wish I could write to my former self, the value of money is at the top of the list of things I'd include.

You can give a cynical, Hansonian explanation of why we don't tell young people enough about the awesomeness of money, and I suppose there'd be some truth to it. But I'm not sure that was my main problem. Growing up, my dad spent a lot of time urging me to go into a high-paying career, to the point giving me advice on what medical specialty to go into. He just didn't do a great job of selling me on it. It wasn't until I learned some economics that I really came to understand why money is so awesome.

(Disclaimer: I don't actually know that much economics, and in fact have never taken an economics course. I just know more than my former self.)

The first thing to understand about money is that the range of things you can get for it is really incredibly huge. Econ bloggers Tyler Cowen and Alex Tabarrok periodically do posts called "markets in everything" where they highlight some of the weirder examples of this, but the weird examples matter less than the obvious examples people just don't think about much. There's a tendency to associate money with a narrow range of things rich people stereotypically spend their money on. Or, in my case growing up in an upper middle-class family, there were the family vacations and boats that my dad seemed to mainly spend his money on, which were nice but didn't seem particularly worth planning my career around.

Yet not only is the range of things you can get with money huge, even with things you can get without money, spending money on them is often a better way of acquiring them. The reason for this is comparative advantage, a concept that gets discussed a lot in the context of nation-states and why free trade is a good idea, but which also works on an individual level. For example, say you're a lawyer who makes $300 an hour, and you're trying to solve the problem of how to keep your house clean. You could spend a couple hours a week doing it yourself—or you could work slightly longer hours and hire someone else to do it for $30 an hour.

The reason this is an example of *comparative* advantage is it doesn't matter if the people you're paying to clean your house are any better at house-cleaning than you. In fact, it works even if they're slightly worse, as long as the difference in house-cleaning ability is overshadowed by the difference in lawyering ability. In econ jargon,

you can have an *absolute* advantage at both lawyering and house-cleaning, and it will still make sense to pay other people to clean your house if they have a *comparative* advantage there. Many people who aren't rich probably assume that when rich people hire other people to do basic tasks for them, it's a frivolous expense, but under the right circumstances it can a matter of economic efficiency.

This point about comparative advantage, when applied to charity, is one of the central insights of the effective altruism movement ("earning to give"). Suppose instead of talking about a lawyer who wants to keep his house clean, we're instead talking about a lawyer who wants to help the local soup kitchen. He could volunteer to help out there in his spare time, but he could also work a little longer hours, donate the money, and enable the soup kitchen to hire more person-hours of work there. Choosing to volunteer rather than give would suggest the lawyer isn't mainly concerned about helping the soup kitchen, but perhaps with warm fuzzies or being seen doing good.

And this doesn't just to small-scale decisions about donating some money vs. volunteering a few hours. It also applies to someone trying to decide between, saying, going into a career in medicine and eventually joining Doctors Without Borders vs. going into a career in finance and using the money you make to pay people to distribute bed nets to stricken regions of the world. (That person was me when I was younger, except the second option wasn't even on my radar.)

Note that while I personally think earning to give is an especially important example of how you can exploit comparative advantage to achieve you're goals, it's also worth emphasizing that it's just a special case of a general principle which can be extremely powerful even if you don't care about making the world a better place.

Given all this, what of the saying "if you want something done right do it yourself"? The answer is that, yes, the difficulty of figuring out who's competent and trustworthy does impose transaction costs on hiring people to do stuff for you, but it's important to remember the costs are finite. When the difference in comparative advantage is large enough, they'll often be worth paying.

Now there are still things money can't buy, at least not literally. But money tends to make them easier to acquire. Take the classic example of happiness: there's a traditional idea (which I've heard attributed to the Greek philosopher Epicurus, though I can't find the source now) that more money makes you happier up to a certain point since it's hard to be happy if you're starving, but beyond that more money doesn't help. It turns out that it's not clear this is actually true—some studies have found more money leads to greater happiness up through the highest income levels examined.

But suppose, in spite of this, that you're an income satisficer, meaning you want to make a certain amount of money and don't care about additional money beyond that. Suppose as long as you have that certain amount of money, you care more about being able to do what you love. And suppose you don't care about being able to make the world a better place through donating to charity. Should you then pursue whatever career you think you'll enjoy the most out of those that pay enough money?

Not necessarily. The way to think about this is to realize that time spent is, in an important sense, an expensive luxury. In economics, there's a concept called opportunity cost, which is closely related to comparative advantage. Opportunity cost asks: by choosing to do something, what's the next-best alternative you're giving up? So for example, by this standard the biggest cost of college for many people will be not tuition, but the time they spent in college that could've been spent working. Even

if you didn't go to summer classes, didn't study all that much, and were only qualified for minimum wage jobs, it still easily adds up to more than the cost of a state school in the US.

A lot of things turn out to be like this: when you translate the cost in time into a monetary value, time is *the* biggest component of the cost. Once you start thinking in those terms, it becomes easier to see that just as there are two ways to convert time into a clean house (clean it yourself, or work at a job where you have the comparative advantage and pay someone else to clean it), there are two ways to maximize the amount of time you spend doing things you enjoy: find a job you mostly enjoy, or else find a high paying job you hate and work part-time / take frequent long sabbaticals / work hard when you're young, then retire early.

People tend not to even consider the second set of options because they've been sold a model of "work nine to five for fifty weeks a year from college graduation until you qualify for Social Security," and you are nudged towards that model somewhat by *employers* assuming it. But it's not mandatory, and if you acquire in-demands skills that can translate into greater flexibility. I have a friend who's a dev consultant who recently took a month sabbatical from her job and then quit entirely without having another one lined up because (1) she makes enough money she doesn't need to work year-round and (2) her skills are sufficiently in-demand that she's not worried about her ability to get another job when she wants one.

On the flip side, to understand one of the main problems with the "get a job you love" strategy, consider the extreme case: a job you'd do for free. The problem with such jobs is that they tend to be jobs other people are willing to do them for free too. That makes it hard for anyone to get paid. For example, I love writing, and I'm doing it for free right now. But it turns out lots of other people feel the same way, and the internet has made it really easy for all of us to distribute our writing for free, and now it's even harder to get paid as a writer than it was during the age of print.

This is just one example, but I suspect there's a *systematic* reason why the "get a job you love" strategy tends to produce outcomes you didn't really want: it can make it harder to see what tradeoffs you're *really* making between money and time spent doing things you want to do for their own sake. In the worst case, you end up getting the worst of both worlds: you become a college professor because you think it will pay okay (if not great), and you'll get to devote all your time to the life of the mind. But you end up adjuncting for what's effectively minimum wage while spending most of your time dealing with undergrads who are just taking the course for the elective and only care about getting an A with as little effort as possible.

I'm not saying *everyone* should optimize solely for money in choosing their career. But at the very least, it's worth putting considerable effort into finding out how much you could (perhaps not immediately, but after but in a year or several) if you did optimize for money. That way, you'll at least know the tradeoff you're making when you chose a different career.

And by the way, if you're reading LessWrong, odds are you're fairly smart, and may be underestimating how monetizable your intelligence is. I'd like to repeat the advice given by other people in the online rationalist community to look into programming as a career choice. I'm currently doing App Academy and highly recommend it, if you do apply tell them I sent you. You may also be able to get good information on choosing a career from 80,000 Hours.

I Will Pay $500 To Anyone Who Can Convince Me To Cancel My Cryonics Subscription

Background:

On the most recent LessWrong readership survey, I assigned a probability of 0.30 on the cryonics question. I had previously been persuaded to sign up for cryonics by reading the sequences, but this thread and particularly this comment lowered my estimate of the chances of cryonics working considerably. Also relevant from the same thread was ciphergoth's comment:

> By and large cryonics critics don't make clear exactly what part of the cryonics argument they mean to target, so it's hard to say exactly whether it covers an area of their expertise, but it's at least plausible to read them as asserting that cryopreserved people are information-theoretically dead, which is not guesswork about future technology and would fall under their area of expertise.

Based on this, I think there's a substantial chance that there's information out there that would convince me that the folks who dismiss cryonics as pseudoscience are essentially correct, that the right answer to the survey question was epsilon. I've seen what seem like convincing objections to cryonics, and it seems possible that an expanded version of those arguments, with full references and replies to pro-cryonics arguments, would convince me. Or someone could just go to the trouble of showing that a large majority of cryobiologists really do think cryopreserved people are information-theoretically dead.

However, it's not clear to me how well worth my time it is to seek out such information. It seems coming up with decisive information would be hard, especially since e.g. ciphergoth has put a lot of energy into trying to figure out what the experts think about cryonics and come away without a clear answer. And part of the reason I signed up for cryonics in the first place is because it doesn't cost me much: the largest component is the life insurance for funding, only $50 / month.

So I've decided to put a bounty on being persuaded to cancel my cryonics subscription. If no one succeeds in convincing me, it costs me nothing, and if someone does succeed in convincing me the cost is less than the cost of being signed up for cryonics for a year. And yes, I'm aware that providing one-sided financial incentives like this requires me to take the fact that I've done this into account when evaluating anti-cryonics arguments, and apply extra scrutiny to them.

Note that there are several issues that ultimately go in to whether you should sign up for cryonics (the neuroscience / evaluation of current technology, estimate of the probability of a "good" future, various philosophical issues), I anticipate the greatest chance of being persuaded from scientific arguments. In particular, I find questions about personal identity and consciousness of uploads made from preserved brains confusing, but think there are *very* few people in the world, if any, who are likely to have much chance of getting me un-confused about those issues. The offer is blind to the exact nature of the arguments given, but I mostly foresee being persuaded by the neuroscience arguments.

And of course, I'm happy to listen to people tell me why the anti-cryonics arguments are wrong and I should stay signed up for cryonics. There's just no prize for doing so.

Literature-review on cognitive effects of modafinil (my bachelor thesis)

Modafinil is probably the most popular cognitive enhancer. LessWrong seems pretty interested in it. The incredible Gwern wrote an excellent and extensive article about it.

Of all the stimulants I tried, modafinil is my favorite one. There are more powerful substances like e.g. amphetamine or methylphenidate, but modafinil has much less negative effects on physical as well as mental health and is far less addictive. All things considered, the cost-benefit-ratio of modafinil is unparalleled.

For those reasons I decided to publish my bachelor thesis on the cognitive effects of modafinil in healthy, non-sleep deprived individuals on LessWrong. Forgive me its shortcomings.

Here are some relevant quotes:

Introduction:

> ...the main research question of this thesis is if and to what extent modafinil has positive effects on cognitive performance (operationalized as performance improvements in a variety of cognitive tests) in healthy, non-sleep deprived individuals.... The abuse liability and adverse effects of modafinil are also discussed. A literature research of all available, randomized, placebo-controlled, double-blind studies which examined those effects was therefore conducted.

Overview of effects in healthy individuals:

> ...Altogether 19 randomized, double-blind, placebo-controlled studies about the effects of modafinil on cognitive functioning in healthy, non sleep-deprived individuals were reviewed. One of them (Randall et al., 2005b) was a retrospect analysis of 2 other studies (Randall et al., 2002 and 2005a), so 18 independent studies remain.
>
> Out of the 19 studies, 14 found performance improvements in at least one of the administered cognitive tests through modafinil in healthy volunteers.
> Modafinil significantly improved performance in 26 out of 102 cognitive tests, but significantly decreased performance in 3 cognitive tests.
>
> ...Several studies suggest that modafinil is only effective in subjects with lower IQ or lower baseline performance (Randall et al., 2005b; Müller et al., 2004; Finke et al., 2010). Significant differences between modafinil and placebo also often only emerge in the most difficult conditions of cognitive tests (Müller et al., 2004; Müller et al., 2012; Winder-Rhodes et al., 2010; Marchant et al., 2009).

Adverse effects:

> ...A study by Wong et al. (1999) of 32 healthy, male volunteers showed that the most frequently observed adverse effects among modafinil subjects were

headache (34%), followed by insomnia, palpitations and anxiety (each occurring in 21% of participants). Adverse events were clearly dose- dependent: 50%, 83%, 100% and 100% of the participants in the 200 mg, 400 mg, 600 mg, and 800 mg dose groups respectively experienced at least one adverse event. According to the authors of this study the maximal safe dosage of modafinil is 600 mg.

Abuse potential:

...Using a randomized, double-blind, placebo-controlled design Rush et al. (2002) examined subjective and behavioral effects of cocaine (100, 200 or 300 mg), modafinil (200, 400 or 600 mg) and placebo in cocaine users....Of note, while subjects taking cocaine were willing to pay $3 for 100 mg, $6 for 200 mg and $10 for 300 mg cocaine, participants on modafinil were willing to pay $2, regardless of the dose. These results suggest that modafinil has a low abuse liability, but the rather small sample size (n=9) limits the validity of this study.

The study by Marchant et al. (2009) which is discussed in more detail in part 2.4.12 found that subjects receiving modafinil were significantly less ($p<0,05$) content than subjects receiving placebo which indicates a low abuse potential of modafinil. In contrast, in a study by Müller et al. (2012) which is also discussed in more detail above, modafinil significantly increased ($p<0,05$) ratings of "task-enjoyment" which may suggest a moderate potential for abuse.

...Overall, these results indicate that although modafinil promotes wakefulness, its effects are distinct from those of more typical stimulants like amphetamine and methylphenidate and more similar to the effects of caffeine which suggests a relatively low abuse liability.

Conclusion:

In healthy individuals modafinil seems to improve cognitive performance, especially on the Stroop Task, stop-signal and serial reaction time tasks and tests of visual memory, working memory, spatial planning ability and sustained attention. However, these cognitive enhancing effects did only emerge in a subset of the reviewed studies. Additionally, significant performance increases may be limited to subjects with low baseline performance. Modafinil also appears to have detrimental effects on mental flexibility.

...The abuse liability of modafinil seems to be small, particularly in comparison with other stimulants such as amphetamine and methylphenidate. Headache and insomnia are the most common adverse effects of modafinil.

...Because several studies suggest that modafinil may only provide substantial beneficial effects to individuals with low baseline performance, ultimately the big question remains if modafinil can really improve the cognitive performance of already high-functioning, healthy individuals. Only in the latter case modafinil can justifiably be called a genuine cognitive enhancer.

You can download the whole thing below. (Just skip the sections on substance-dependent individuals and patients with dementia. My professor wanted them.)

Effects of modafinil on cognitive performance in healthy individuals, substance-dependent individuals and patients with dementia