Best of LessWrong: October 2013

1. A Voting Puzzle, Some Political Science, and a Nerd Failure Mode
2. How to Become a 1000 Year Old Vampire
3. Mental Context for Model Theory
4. How habits work and how you may control them
5. Bayesianism for Humans
6. [Link] Low-Hanging Poop
7. How to Learn from Experts
8. What Can We Learn About Human Psychology from Christian Apologetics?
9. Only You Can Prevent Your Mind From Getting Killed By Politics
10. Systematic Lucky Breaks
11. Does Goal Setting Work?
12. Meditation Trains Metacognition

Best of LessWrong: October 2013

A Voting Puzzle, Some Political Science, and a Nerd Failure Mode

In grade school, I read a series of books titled *Sideways Stories from Wayside School* by Louis Sachar, who you may know as the author of the novel *Holes* which was made into a movie in 2003. The series included two books of math problems, *Sideways Arithmetic from Wayside School* and *More Sideways Arithmetic from Wayside School,* the latter of which included the following problem (paraphrased):

The students have Mrs. Jewl's class have been given the privilege of voting on the height of the school's new flagpole. She has each of them write down what they think would be the best hight for the flagpole. The votes are distributed as follows:

- 1 student votes for 6 feet.
- 1 student votes for 10 feet.
- 7 students vote for 25 feet.
- 1 student votes for 30 feet.
- 2 students vote for 50 feet.
- 2 students vote for 60 feet.
- 1 student votes for 65 feet.
- 3 students vote for 75 feet.
- 1 student votes for 80 feet, 6 inches.
- 4 students vote for 85 feet.
- 1 student votes for 91 feet.
- 5 students vote for 100 feet.

At first, Mrs. Jewls declares 25 feet the winning answer, but one of the students who voted for 100 feet convinces her there should be a runoff between 25 feet and 100 feet. In the runoff, each student votes for the height closest to their original answer. But after that round of voting, one of the students who voted for 85 feet wants *their* turn, so 85 feet goes up against the winner of the *previous* round of voting, and the students vote the same way, with each student voting for the height closest to their original answer. Then the same thing happens again with the 50 foot option. And so on, with each number, again and again, "very much like a game of tether ball."

Question: if this process continues until it settles on an answer that can't be beaten by any other answer, how tall will the new flagpole be?

Answer (rot13'd): fvkgl-svir srrg, orpnhfr gung'f gur zrqvna inyhr bs gur bevtvany frg bs ibgrf. Naq abj lbh xabj gur fgbel bs zl svefg rapbhagre jvgu gur zrqvna ibgre gurberz.

Why am I telling you this? There's a minor reason and a major reason. The minor reason is that this shows it is possible to explain little-known academic concepts, at least certain ones, in a way that grade schoolers will understand. It's a data point that fits nicely with what Eliezer has written about how to explain things. The major reason, though, is that a month ago I finished my systematic read-through of the sequences and while I generally agree that they're awesome (perhaps moreso than most people; I didn't see the problem with the metaethics sequence), I thought the mini-discussion of political parties and voting was on reflection weak and indicative of a broader nerd failure mode.

TLDR (courtesy of lavalamp):

1. Politicians probably conform to the median voter's views.
2. Most voters are not the median, so most people usually dislike the winning politicians.
3. But people dislike the politicians for different reasons.
4. Nerds should avoid giving advice that boils down to "behave optimally". Instead, analyze the reasons for the current failure to behave optimally and give more targeted advice.

Advance warning for heavy US slant, at least in terms of examples, though the theory is applicable everywhere.

The median voter theorem

The median voter theorem was first laid out in a paper by Duncan Black titled "On the Rationale of Group Decision-Making," which imagine's a situation very much like Mrs. Jewls' class voting on the flagpole height: a committee passes a motion by majority vote, and then it considers various motions to amend the original motion, each of which itself needs a simple majority to pass. Each member of the committee has preferences over the range of possible motions, and furthermore:

> While a member's preference curve may be of any shape whatever, there is reason to expect that, in some important practical problems, the valuations actually carried out will tend to take the form of isolated points on single-peaked curves. This would be particularly likely to happen were the committee considering different possible sizes of a numerical quantity and choosing one size in preference to the others. It might be reaching a decision, say, with regard to the price of a product to be marketed by a firm, or the output for a future period, or the wage rate of labor, or the height of a particular tax, or the legal school-leaving age, and so on.

Or, for that matter, the height of a flagpole. Black shows that on his assumptions, the committee will eventually settle on the version of the motion favored by the median committee member.

Again, you may be asking, so what? Most people don't care about understanding the behavior of committees, especially not compared to their passion for national presidential elections. And elections for political office don't use a tether ball-like system of having head-to-head matchup after head-to-head matchup until you've finally found the candidate the median voter wants. There's one election with two (or if you're lucky, three) major candidates and that's it.

The relevance to electoral politics comes in when you allow for the possibility of candidates shaping themselves and their platforms to appeal to the median voter. The candidate who does this should be invincible - at least, until the other candidate does the same thing, at which point the election becomes a closer call. The idea of candidates shaping themselves to voter preferences is key; I started off this post with the flagpole example partly to emphasize that. And there are other assumptions you have to make to get to the conclusion that candidates will *actually* behave this way.

But before we get in to that, let's compare the median voter picture to the picture Eliezer put forward in the posts linked above:

> Forget that Congresspeople on both sides of the "divide" are more likely to be lawyers than truck drivers. Forget that in training and in daily life, they have far more in common with each other than they do with a randomly selected US citizen from their own party. Forget that they are more likely to hang out at each other's expensive hotel rooms than drop by your own house. Is there a political divide - a divide of policies and interests - between Professional Politicians on the one hand, and Voters on the other?
>
> Well, let me put it this way. Suppose that you happen to be socially liberal, fiscally conservative. Who would you vote for?
>
> Or simplify it further: Suppose that you're a voter who prefers a smaller, less expensive government - should you vote Republican or Democratic? Or, lest I be accused of color favoritism, suppose that your voter preference is to get US troops out of Iraq. Should you vote Democratic or Republican?
>
> One needs to be careful, at this point, to keep track of the distinction between marketing materials and historical records. I'm not asking which political party stands for the idea of smaller government - which football team has "Go go smaller government! Go go go!" as one of its cheers. (Or "Troops out of Iraq! Yay!") Rather, over the last several decades, among Republican politicians and Democratic politicians, which group of Professional Politicians shrunk the government while it was in power?
>
> And by "shrunk" I mean "shrunk". If you're suckered into an angry, shouting fight over whether Your Politicians or Their Politicians grew the government slightly less slowly, it means you're not seeing the divide between Politicians and Voters. There isn't a grand conspiracy to expand the government, but there's an incentive for each individual politician to send pork to campaign contributors, or borrow today against tomorrow's income. And that creates a divide between the Politicians and the Voters, as a class, for reasons that have nothing to do with colors and slogans.

Eliezer observes that there doesn't seem to be much difference between the two parties, and concludes that they are colluding (albeit probably not by explicit agreement) to advance their own interests at the expense of the voters'. The politicians don't offer the voters any real choice, but get voters to vote for them anyway though misleading party labels and the argument that, if they don't vote for a major-party candidate, they're "throwing their vote away."

However, the observation that there doesn't seem to be much difference between the two parties can *also* be explained by the hypothesis that politicians are shaping themselves to appeal to the median voter. This fact alone doesn't show that the median voter model is *right...* but it does show that the mere fact of there not being much difference between the two parties doesn't show the "colluding politicians" model is right *either.*

So how well does the median voter theorem capture reality? One problem for the model is that it potentially breaks down if the choices don't fit onto a nice, linear spectrum. Suppose, for the sake of a simplified example, that only three people vote in a particular presidential election. Suppose, furthermore, that the three voters have the following set of preferences:

- Alice prefers Obama to Romney, and Romney to Ron Paul
- Bob prefers Romney to Ron Paul, and Ron Paul to Obama
- Carol prefers Ron Paul to Obama, and Obama to Romney

Given this set of voters and their preferences, in an Obama vs. Romney contest, Obama will win; in a Romney vs. Ron Paul contest, Romney will win; but in a Ron Paul vs. Obama contest, Ron Paul will win.

However, the median voter theorem seems to be a pretty good model in practice in spite of such problems. Roger Congleton, in an article in the *Encyclopedia of Public Choice*, writes:

> Although theoretical arguments suggest that the applicability of the median voter model may be very limited, the empirical evidence suggests otherwise. There is a large body of evidence that suggests median voter preferences over policies are (largely) of the sort which can be mapped into a single issue space while retaining "single peakedness" Poole and Daniels (1985) find that 80-90% of all the recorded votes in the US Congress can be explained with a one dimensional policy space. Stratmann (1996) finds little evidence of cycling across Congressional votes over district specific grants.
>
> Moreover, the median voter model has a very good empirical track record in public finance as a model of fiscal policy across states and through time. Recent studies show that the median voter model can explain federal, state, and local spending, as well as international tariff policies. Congleton and Shughart (1990) Congleton and Bennett (1995) suggest that the median voter model provides a better explanation of large scale public programs than comparable interest group models. This is not to suggest that the median voter always exercises the same degree of control over public policy irrespective of political institutions. Holcombe (1980) and Frey (1994) report significant policy difference between representative and direct forms of democracy that would not exist unless significant agency problems exist within representative government. Moreover, statistical tests can never prove that a particular model is correct, only that it is more likely to be correct than false. However, in general, the median voter model appears to be quite robust as a model of public policy formation in areas where the median voter can credibly be thought to understand and care about public policy.
>
> The empirical evidence suggests that the median voter model can serve as a very useful first approximation of governance within democratic polities. As a consequence, the median voter model continues to function as an analytical point of departure for more elaborate models of policy formation within democracies in much the same way that the competitive model serves the micro economics literature.

In the American political system, the effect of the median voter theorem is blunted somewhat by the primary system. It's a commonplace among American political commentators that politicians must appeal to the "base" during the primaries, then swing towards the center for the general election. Of course, politicians can't suddenly become *perfectly* centrist once they secure their party's nomination; their swing towards the center has to be done in a way that's at least superficially consistent with their previous pandering to their base. These observations suggest that, while reality doesn't perfectly match the idealized model, there's still a lot of truth to it.

(Note: a site search for previous discussion of the median voter theorem on LessWrong turned up a comment by Carl Shulman that mentioned "the need to motivate one's base to vote/volunteer/contribute the ideological lumpiness" as probably having an effect similar to the effect of primaries. I wouldn't have thought they were as important as primaries but I can believe Carl here.)

The tendency of politicians to position themselves wherever the center of public opinion is currently at can be striking on specific issues. For example, public support for gay rights has increased greatly in the past two decades. In that time period, positions which once got Bill Clinton demonized by the religious right as an agent of the homosexual agenda (like Don't Ask Don't Tell) became the "conservative" position. Progress, but in terms of the public stances of politicians, it's progress that came not in the form of dramatic shifts but cautious adjustments.

By the time of the 2008 campaign, Republican nominee John McCain was voicing vague support for "legal agreements" between same sex couples, while rejecting same-sex marriage. At the same time, he suggested the issue could be punted to the states. Meanwhile, Obama's position was only slightly more liberal: clearer support for civil unions (but again not full marriage equality), and similar suggestions that the issue could be left to the states.

Four years later in 2012, Obama *finally* mentioned in an interview that he'd changed his mind and now supported same-sex marriage. By that time, figures from Rick Santorum to Rick Warren to Sarah Palin had begun telling the press that they, too, have gay friends. Since that time, the Obama administration has only taken modest concrete steps to support gay marriage: a narrowly-worded brief opposing California's Proposition 8, a decision not to defend the Defense of Marriage Act in court, and that's about it.

From the point of view of the median voter model, the way to explain both the behavior of liberals like Obama and conservatives like McCain and Santorum is that both groups are trying to avoid straying very far from the position of the median voter, so as to not alienate them and lose their vote. It's significant that in 2008, the polls showed that public opinion was roughly divided into thirds on gay marriage, with about a third totally opposed, a third supporting civil unions, and a third supporting full marriage equality. Obama's announcement that he supported gay marriage came *after* numerous polls showed 50-some percent of Americans supporting gay marriage.

Some readers may be wondering how this analysis fits with the current polarization in Congress. The answer is, "perfectly." The median voter theorem leads us to expect that politicians running against each other should adopt similar views, but even in its most idealized form, it says nothing about members of the same legislature should have similar views. In fact, it *predicts* polarized legislatures in situations where (1) members of the legislature are elected by geographic region and (2) the electorate itself is polarized by geographic region.

This is what we see in the US, where a big-city congressional district can be *much* more liberal than a rural one. Many members of the House of Representatives probably have *more* to worry about from a more-extreme primary challenger within their own party than from a general election challenger from the other party. Caveat: I've tried looking up data on the voting records of various House members, and while there's clearly a correlation between the tendencies of their respective districts, the correlation is not as strong as I expected. I'd be curious to hear if anyone out there knows more about this issue of polarization and geography.

Voting systems, voting strategies, and knowing your fellow voters

So elections in the US may not offer voters much choice, but that's better explained by the median voter theorem than by politicians colluding against voters. Political science also provides a second objection to Eliezer's analysis of the two-party system in America: Duverger's Law, which says that in a system like ours (where everyone votes for one candidate and whoever gets the most votes wins), the system will tend to converge on having two main political parties, due to standard reasoning about not throwing your vote away. A corollary is that you can get a multiparty system by using proportional representation, which is used in many countries around the world including Spain, Portugal, Italy, Germany, and Israel.

There are some apparent exceptions to Duverger's Law, such as Canada, which has long had a multiparty system in spite of using a voting system similar to that of the US. However, a friend of mine who follows Canadian politics tells me that what really happens in Canada isn't *that* far from what you would expect given Duverger's Law. Currently, the three largest parties are the Conservative Party, the New Democratic Party (NDP), and the Liberal Party. It used to be that the NDP was a relatively small party with positions well to the left of the Liberals, but this is no longer true. Instead of offering Canadian voters two different flavors of liberalism, the current situation is that in any given election for any given seat in parliament, the NDP candidate and the Liberal candidate put a lot of effort into arguing over who has the best chance of beating the Conservatives.[1]

So suppose you're an American or Canadian or British voter, looking at the major-party candidates in the next election, and finding that *none* of them are a good fit for your political views, what you should conclude? First, given that the median voter theorem is a pretty good model of how elections actually work, you should probably take your as evidence that your views are a good ways away from those of the median voter. And if the views of the voters are sufficiently varied, *a majority of voters could find themselves in the same position as you.*

In the flagpole problem at the start of this post, the only one student originally wanted the height that ends up winning. Actually, there's a subtle joke I left out of my paraphrase: the student who wanted 65 feet was Kathy, who elsewhere in the series was established as hating everyone and loving to see bad things happen. Or, to use the gay marriage example: in the 2008 election, the ~1/3 of voters who supported gay marriage didn't have a major party candidate who supported their views (and voters totally opposed to gay marriage and civil unions may not have been terribly happy with their choices either).

To throw off the yoke of the existing major parties, it isn't enough for most voters to reject their platforms. They need to reject their platforms in more or less the same direction. In "Stop Voting for Nincompoops," Eliezer mentions having anti-interventionist foreign policy views, and based on that, maybe he would say Obama is a nincompoop for being too interventionist, too willing to kill foreigners in the name of fighting terrorism. If so, I'd be sympathetic. But even if a majority of Americans *agreed* that Obama is a nincompoop, it wouldn't follow that they *agree he is a nincompoop for being too willing to kill foreigners in the name of fighting terrorism.* Many of them probably think he's a nincompoop for not doing nearly enough to fight terrorism, and maybe even being secretly on the side of the terrorists.[2]

That's because median voter analysis suggests that if none of the main candidates in an election are a good fit for your views, this is a sign that your views are a good ways from those of the median voter, and as a corollary there must be people out there whose views differ from the median voter's in the *opposite* direction, and therefore

would seem even *more* repugnant to you. (Never forget that half the population is below average.)

In "Stop Voting for Nincompoops," Eliezer quotes from Douglas Adams' novel *So Long And Thanks For All The Fish:*

> "The leaders are lizards. The people hate the lizards and the lizards rule the people."
>
> "Odd," said Arthur, "I thought you said it was a democracy."
>
> "I did," said Ford, "It is."
>
> "So," said Arthur, hoping he wasn't sounding ridiculously obtuse, "why don't the people get rid of the lizards?"
>
> "It honestly doesn't occur to them," said Ford. "They've all got the vote, so they all pretty much assume that the government they've voted in more or less approximates to the government they want."
>
> "You mean they actually vote for the lizards?"
>
> "Oh yes," said Ford with a shrug, "of course."
>
> "But," said Arthur, going for the big one again, "why?"
>
> "Because if they didn't vote for a lizard," said Ford, "the wrong lizard might get in. Got any gin?"

In light of all the above, let me suggest a modified allegory: the people hate the lizards, and have thought of getting rid of them, but there's disagreement about what to do *after* getting rid of the lizards. Many people favor self-rule, but a very nearly equal number of people favor replacing the lizards with the Demon Acolytes of Yog-Sothoth. Since a few people actually like the lizards, and almost everyone agrees lizards are better than what those *other* people want, lizards are what they get.

Of course, since very few people consider themselves villains, to make the story as realistic as possible, we should imagine that the partisans of Demon Acolytes believe the demons are actually Angels of the Light, and that anyone prideful enough to think autonomy is better than being ruled by angels must be profoundly wicked. Either way, the point is that widespread dislike of the current political situation does not imply widespread support for any particular alternative.

Moving back to the real world again, here's an explanation for US foreign policy under both Bush II and Obama, which I suspect Eliezer would think too cynical, but which I'll mention anyway: maybe the reason the US government is so quick to kill foreigners in the name of fighting terrorism is because the median voter fears terrorism more than they care about the lives of foreigners. I suppose you could argue it isn't so, and the real reason is the median voter doesn't know what impact US foreign policy has on foreigners, but if they cared to know, couldn't they start paying less attention to CNN and more to Al-Jazeera?

Given all this, how should you vote? Well, you shouldn't vote for a third party candidate *because* you think a lot of our problems could be solved if everyone just simultaneously resolved to never vote for (anyone they believed to be) a nincompoop.

If somehow you actually manage to persuade people to everyone to adopt that policy, don't be surprised if disagreements about who the nincompoops are result in nothing really changing, or worse result in a bunch bizarre elections decided by small pluralities.

Beyond that though, I'm not actually sure what the proper strategy is. In spite of everything I've said, maybe the "vote third party to send a message" argument is (sometimes) right. Or maybe there's something to be said for the argument that your vote isn't going to make a difference anyway so you may as well do whatever makes you feel good. So far in my relatively short time as a voter, I've adopted a mixed approach, protest-voting in my two presidential elections but voting for major-party candidates otherwise. But I'm honestly not sure what I'll do in the future. Maybe a seemingly-infinitesimal chance of affecting the election outcome is worth it.

That is not a very exciting way to end an essay this long. Which is why I'm happy to report that that is not how I'm ending this essay, and in fact have been building up to a different general point.

A nerd failure mode regarding human affairs

So at last, I'm ready to explain what I think the broader nerd failure mode here is: they have a tendency to notice that people are failing to behave optimally and then propose, as a solution to this problem, that people switch to behaving optimally.

This is related to, if not quite the same as, the problem Randall Munroe pokes at here. The problem is that if you don't first make a serious effort to figure out *why* people are failing to behave optimally, that can get in the way of figuring out what a better course of action would be. And it makes it almost impossible to figure out how to get people to *actually follow* the better course of action.

If the reason people elect bad leaders is that half the people have views even crazier than those of the leaders they elect, you will not make much progress changing things if you think the problem is a two-party conspiracy against the voters. Or, if you to get people to stop voting for nincompoops, convincing them they should never vote for nincompoops may give you a very different result than you were expecting if they have different ideas from you about who the nincompoops are and what it is about them that qualifies them as nincompoops.

Many readers of LessWrong will have heard of Chesterton's fence already, but let me quote Chesterton's original words at somewhat greater length than is usual:

> In the matter of reforming things, as distinct from deforming them, there is one plain and simple principle; a principle which will probably be called a paradox. There exists in such a case a certain institution or law; let us say for the sake of simplicity, a fence or gate erected across a road. The more modern type of reformer goes gaily up to it and says, "I don't see the use of this; let us clear it away." To which the more intelligent type of reformer will do well to answer: "If you don't see the use of it, I certainly won't let you clear it away. Go away and think. Then, when you can come back and tell me that you do see the use of it, I may allow you to destroy it."
>
> This paradox rests on the most elementary common sense. The gate or fence did not grow there. It was not set up by somnambulists who built it in their sleep. It

> is highly improbable that it was put there by escaped lunatics who were for some reason loose in the street. Some person had some reason for thinking it would be a good thing for somebody. And until we know what the reason was, we really cannot judge whether the reason was reasonable. It is extremely probable that we have overlooked some whole aspect of the question, if something set up by human beings like ourselves seems to be entirely meaningless and mysterious. There are reformers who get over this difficulty by assuming that all their fathers were fools; but if that be so, we can only say that folly appears to be a hereditary disease. But the truth is that nobody has any business to destroy a social institution until he has really seen it as an historical institution. If he knows how it arose, and what purposes it was supposed to serve, he may really be able to say that they were bad purposes, or that they have since become bad purposes, or that they are purposes which are no longer served. But if he simply stares at the thing as a senseless monstrosity that has somehow sprung up in his path, it is he and not the traditionalist who is suffering from an illusion.

In spite of being a conservative Catholic apologist, what Chesterton is saying here isn't crazy. Certainly it helps to know what people's reasons for something were before trying to judge whether they were good ones. I wouldn't go quite as far as Chesterton, since sometimes there's such good evidence something's a bad idea that you can reject it without knowing what people were originally thinking.

But even on much weaker assumptions than Chesterton's, something in the vicinity turns out to be good advice. Even if the fence *was* built by lunatics, that's worth knowing. It's especially worth knowing whether they're still out there, and whether they're likely to try to rebuild the fence after it's been taken down. If they are likely to try that, you need to know so they can be recaptured before taking the fence down, so that the lunatics don't just rebuild it, making the taking-down a waste of effort.

Notes

1. Someone might read this and conclude that, since the two-party system is so awful, and Duverger's Law implies it's a necessary result of our voting system, shouldn't we switch voting systems to something like proportional representation? I'm willing to believe that other systems might be *slightly* better than what we have in the US. Countries that use proportional representation tend to have higher voter turnout, though it's unclear whether the one causes the other. But does anyone think that proportional representation and more major parties makes, say, Germany's government *that much* better than the UK's? For more on voting systems, see Yvain's summary of why no voting system is perfect.

2. Some people reading this might be skeptical of the idea many people would believe something as crazy-sounding as "Obama is secretly on the side of the terrorists." While I think we should be careful about phantom lizardmen and partisan media selectively reporting on the other side's crazies to gin up outrage, sadly, from what I can tell there genuinely are a large number of people out there who believe such right-wing conspiracy theories about Obama. I'm not trying to make a partisan point here, and say this with full awareness of things like 9/11 conspiracy theories on the left.

   Remember, first, that hardly any of us come into contact with a random sampling of our fellow voters on a daily basis. Furthermore, I grew up in a

smallish (pop. ~60k), conservative-leaning town, and occasionally people I barely interacted with in high school will friend me on Facebook, I'll accept because why not, and then I'll start getting their thoughts on politics in my Facebook feed. That may give me a somewhat clearer perspective on this than the averge resident of a liberal big city. I remember when the NSA scandal broke and one girl posted a status update which, while containing civil-libertarian thoughts that I approved of, also contained references to Obama being an illegal president (because, as far as I could tell, birtherism), as well as a reference to Obama's "terroristic ways," whatever that means.

How to Become a 1000 Year Old Vampire

This is based on a concept we developed at the Vancouver Rationalists meetup.

Different experiences level a person up at different rates. You could work some boring job all your life and be 60 and not be much more awesome than your average teenager. On the other hand, some people have such varied and so much life experience that by 30 they are as awesome as a 1000 year old vampire.

This reminds me that it's possible to conduct your life with more or less efficiency, sometimes by orders of magnitude. Further, while we don't have actual life extension, it's content we care about, not run time. If you can change your habits such that you get 3 times as much done, that's like *tripling your effective lifespan*.

So how might one get a 100x speedup and become like a 1000 year old vampire in 10 years? This is absurdly ambitious, but we can try:

Do Hard Things

Some experiences catapult you forward in personal development. You can probably systematically collect these to build formidability as fast as possible.

Paul Graham says that many of the founders he sees (as head of YC) become much more awesome very quickly as need forces them to. This seems plausible and it seems back up by other sources as well. Basically "learn to swim by jumping in the deep end"; people have a tendency to take the easy way that results in less development when given the chance, so the chance to slack off being removed can be beneficial.

That has definitely been my personal experience as well. At work, the head engineer got brain cancer and I got de-facto promoted to head of two of the projects, which I then leveled up to be able to do. It felt pretty scary at first, but now I'm bored and wishing something further would challenge me. (addendum: not bored right now at all; crazy crunch time for the other team, which which I am helping) It seems really hard to just do better without such forcing; as far as I can tell I *could* work much harder than now, but willpower basically doesn't exist so I don't.

On that note, a friend of mine got big results from joining the Army and getting tear gassed in a trench while wet, cold, exhausted, sleep deprived, and hungry, which pushed him through stuff he wouldn't have thought he could deal with. Apparently it sortof re-calibrated his feelings about how well he should be doing and how hard things are such that he is now a millionaire and awesome.

So the mechanism behind a lot of this seems to be recalibrating what seems hard or scary or beyond your normal sphere. I used to be afraid of phone calls and doing weird stuff like climbing trees in front of strangers, but not so much anymore; it feels like I just forget that they were scary. In the case of the phone there were a few times where I didn't have time to be scared, I needed to just get things done. In the case of

climbing trees, I did it on my own enough for it to become normalized so that it didn't even come up that people would see me, because it didn't seem weird.

So tying that back in, there are experiences that you can put yourself into to force that normalization and acclimatization to hard stuff. For example, some people do this thing called "Rejection Therapy" or "Comfort Zone Expansion", basically going out and doing embarrassing or scary things deliberately to recalibrate your intuitions and teach your brain that they are not so scary.

On the failure end, self-improvement projects tend to fail when they require constant application of willpower. It's just a fact that you will fall off the wagon on those things. So you have to make it impossible to fall off the wagon. You have to make it scarier to fall off the wagon than it is to level up and just do it. This is the idea behind Beeminder, which takes your money if you don't do what your last-week self said you would.

I guess the thesis behind all this is that these level-ups are permanent, in that they make you more like a 1000 year old vampire, and you don't just go back to being your boring old mortal self. If this is true, the implication that you should seek out hard stuff seems pretty interesting and important.

Broadness of Experience

Think of a 1000 year old vampire; they would have done everything. Fought in battles, led armies, built great works, been in love, been everywhere, observed most aspects of the human experience, and generally seen it all.

Things you can do have sharply diminishing returns; the first few times you watch great movies is most of the benefit thereof, likewise with video games, 4chan, most jobs, and most experiences in general. Thus it's really important to switch around the things you do a lot so that you stay in that sharp initially growing part of the learning curve. You can get 90% of the vampire's experience with 10% of his time investment if you focus on those most enlightening parts of each experience.

So besides doing hard things that level you up, you can get big gains by doing *many* things and switching as soon as you get bored (which is hopefully calibrated to how challenged you are).

You may remember early in the Arabian revolutions in Libya, an American student took the summer off college to fight in the revolution. I bet he learned a lot. If you could do enough things like that, you'd be well on your way to matching the vampire.

This actually goes hand in hand with doing hard things; when you're not feeling challenged (you're on the flat part of that experience curve), its probably best to throw yourself face first into some new project, both because it's new, and because it's hard.

Switching often has the additional benefit of normalizing strategic changes and practicing "what should I be doing"-type thoughts, which can't hurt if you intend to actually do useful stuff with your life.

There are probably many cases where full on switching is not best. For example, you don't become an expert in X by switching out of X as soon as you know the basics. It might be that you want to switch often on side-things but go deep on X. Alternatively,

you probably want to do some kind of switch every now and then in X, maybe look at things from a different perspective, tackle a different problem, or something like that. This is the Deliberate Practice theory of expertise.

So don't forget the shape of that experience curve. As soon as you start to feel that leveling off, find a way to make it fresh again.

Do Things Quickly

Another big angle on this idea is that every hour is an opportunity, and you want to make the best of them. This seems totally obvious but I definitely "get it" a lot more having thought about it in terms of becoming a 1000 year old vampire.

A big example is procrastination. I have a lot of things that have been hanging around on my todo list for a long time, basically oppressing me by their presence. I can't relax and look to new things to do while there's still that one stupid thing on my todo list. The key insight is that if you process the stuff on your todo list *now* instead of slacking now and doing it later, you get it out of the way and then you can do something *else* later, and thereby become a 1000 year old vampire faster.

So a friend and I have internalized this a bit more and started really noticing those opportunity costs, and actually started knocking things off faster. I'm sure there's more where that came from; we are nowhere near optimal in Doing It Now, so it's probably good to meditate on this more.

As a concrete example, I'm writing tonight because I realized that I need to just get all my writing ideas out of the way to make room for more awesomeness.

The flipside of this idea is that a lot of things are complete wastes of time, in the sense that they just burn up lifespan and don't get you anything, or even weaken you.

Bad habits like reading crap on the Internet, watching TV, watching porn, playing video games, sleeping in, and so on are obvious losses. It's really hard to internalize that, but this 1000-year-old-vampire concept has been helpful for me by making the magnitude of the cost more salient. Do you want to wake up when you're 30 and realize you wasted your youth on meaningless crap, or do you want to get off your ass and write that thing you've been meaning to *right now, and be a fscking vampire in 10 years*?

It's not just bad habits, though; a lot of it is your broader position in life that wastes time or doesn't. For example, repetitive wage work that doesn't challenge you is really just trading a huge chunk of your life for not even much money. Obviously sometimes you have to, but you have to realize that trading away half your life is a pretty raw deal that is to be avoided. You don't even really *get* anything for commuting and housework. Maybe I really should quit my job soon...

I have 168 hours a week, of which only 110 are feasible to use (sleep), and by the time we include all the chores, wage-work, bad habits, and procrastination, I probably only *live* 30 hours a week. That's bullshit; three quarters of my life pissed away. I could live *four* times as much if I could cut out that stuff.

So this is just the concept of time opportunity costs dressed up to be more salient. Basic economics concepts seem really quite valuable in this way.

Do it now so you can do something else later. Avoid crap work.

Social Environment and Stimulation

I notice that I'm most alive and do my best intellectual work when talking to other people who are smart and interested in having deep technical conversations. Other things like certain patterns of time pressure create this effect where I work many times harder and more effectively than otherwise. A great example is technical exams; I can blast out answers to hundreds of technical questions at quite a rate.

It seems like a good idea to induce this state where you are more alive (is it the "flow" state?) if you want to live more life. It also seems totally possible to do so more often by hanging out with the right people and exposing yourself to the right working conditions and whatnot.

One thing that will come up is that it's quite draining, in that I sometimes feel exhausted and can't get much done after a day of more intense work. Is this a real thing? Probably. Still, I'm nowhere near the limit even given the need to rest, in general.

I ought to do some research to learn more about this. If it's connected to "flow", there's been a lot of research, AFAIK.

I also ought to just hurry up and move to California where there is a proper intellectual community that will stimulate me much better than the meager group of brains I could scrape together in Vancouver.

The other benefit of a good intellectual community is that they can incentivize doing cooler things. When all your friends are starting companies or otherwise doing great work, sitting around on the couch *feels* like a really bad idea.

So if we want to live more life, finding more ways to enter that stimulated flow state seems like a prudent thing to do, whether that means just making way for it in your work habits, putting yourself in more challenging social and intellectual environments, or whatever.

Adding It Up

So how fast can we go overall if we do all of this?

By seeking many new experiences to keep learning, I think we can plausibly get 10x speedup over what you might do by default. Obviously this can be more or less, based on circumstances and things I'm not thinking of.

On top of that, it seems like I could do 4x as much by maintaining a habit of doing it *now* and avoiding crap work. How to do this, I don't know, but it's possible.

I don't know how to estimate the actual gains from a stimulating environment. It seems like it could be really really high, or just another incremental gain in efficiency, depending how it goes down. Let's say that *on top* of the other things, we can realistically push ourselves 2x or 3x harder by social and environmental effects.

Doing hard things seems huge, but also quite related to the doing new things angle that we already accounted for. So explicitly remembering to do hard things on top of that? Maybe 5x? This again will vary a lot based on what opportunities you are able to find, and unknown factors, but 5x seems safe enough given mortal levels of ingenuity and willpower.

So all together, someone who:

- Often thinks about where they are on the experience curve for everything they do, and takes action on that when appropriate,
- Maintains a habit of doing stuff *now* and visualizing those opportunity costs,
- Puts themselves in a stimulating environment like the bay area intellectual community and surrounds themselves with stimulating people and events,
- Seeks out the hardest character-building experiences like getting tear gassed in a trench or building a company from scratch,

Can plausibly get 500x speedup and live 1000 normal years in 2. That seems pretty wild, but none of these things are particularly out there, and people like Elon Musk or Eliezer Yudkowsky do seem to do around that magnitude more than the average joe.

Perhaps they don't multiply quite that conveniently, or there's some other gotcha, but the target seems reachable, and these things will help. On the other hand, they almost certainly self-reinforce; a 1000 year old vampire would have mastered the art of living life life at ever higher efficiencies.

This does seem to be congruent with all this stuff being power-law distributed, which of course makes it difficult to summarize by a single number like 500.

The final question of course is what *real* speedup we can expect you or I to gain from writing or reading this. Getting more than 2 or 3 times by having a low-level insight or reading a blog post seems stretching of the imagination, never mind 500 times. But still, power laws happen. There's probably massive payoff to taking this idea seriously.

Mental Context for Model Theory

I'm reviewing the books on the MIRI course list. After my first four book reviews I took a week off, followed up on some dangling questions, and upkept other side projects. Then I dove into *Model Theory*, by Chang and Keisler.

It has been three weeks. I have gained a decent foundation in model theory (by my own assessment), but I have not come close to completing the textbook. There are a number of other topics I want to touch upon before December, so I'm putting *Model Theory* aside for now. I'll be revisiting it in either January or March to finish the job.

In the meantime, I do not have a complete book review for you. Instead, this is the first of three posts on my experience with model theory thus far.

This post will give you some framing and context for model theory. I had to hop a number of conceptual hurdles before model theory started making sense — this post will contain some pointers that I wish I'd had three weeks ago. These tips and realizations are somewhat general to learning any logic or math; hopefully some of you will find them useful.

Shortly, I'll post a summary of what I've learned so far. For the casual reader, this may help demystify some heavily advanced parts of the Heavily Advanced Epistemology sequence (if you find it mysterious), and it may shed some light on some of the recent MIRI papers. On a personal note, there's a lot I want to write down & solidify before moving on.

In follow-up post, I'll discuss my experience struggling to learn something difficult on my own — model theory has required significantly more cognitive effort than did the previous textbooks.

Between what was meant and what was said

Model theory is an abstract branch of mathematical logic, which itself is already too abstract for most. So allow me to motivate model theory a bit.

At its core, model theory is the study of what you *said*, as opposed to what you *meant*. To give some intuition for this, I'll re-tell an overtold story about an ancient branch of math.

In olden times, Euclid built Geometry upon five axioms:

1. You can draw a straight line segment between two points.
2. You can extend line segments into infinitely straight lines.
3. You can draw a circle from a straight line segment, with the center at one end and radius the line segment.
4. All right angles are congruent.
5. If two lines are drawn which intersect a third in such a way that the sum of the inner angles on one side is less than two right angles, then the two lines inevitably must intersect each other on that side if extended far enough.

One of these things is not like the other. The fifth axiom is the only one which requires some effort to understand. Intuitively, it states that parallel lines do not intersect. This

statement irked Euclid for reasons apart from the ugliness of the axiom.

The fact that parallel lines do not intersect seems like it should follow from the definition of lines and angles. It doesn't seem like something we should have to specify *in addition*. That we must *assume* parallel lines do not intersect (rather than *proving* it) was long seen as a wart on geometry.

This wart irked mathematicians for millennia, until finally it was discovered that the fifth axiom is independent of the other four. You can build consistent systems where parallel lines intersect. You can build consistent systems where they diverge.

This seemed crazy, at the time: parallel straight lines cannot diverge! Surely, a geometry in which they do is absurd!

The problem is that mathematicians were imagining "straight lines" in their head that did not match the mathematical objects specified by the first four axioms of Euclid.

This mistake was invited by names which Euclid chose. "Straight lines" invoke a mental image that is more specific than that which the axioms describe. If you detach the provocative words from the axioms

1. You can make a `LUME` between any two `PTARS`
2. You can extend a `LUME` into a `SLUME`
3. ...

and so on, then it's much easier to understand that the `LUME`s which Euclid's axioms describe may not match up with the image of a "straight line" in your head. It is much easier to understand that there may be interpretations of `LUME` which do not obey the fifth postulate.

In fact, if you take Euclid's first four postulates, there are many possible interpretations in which "straight line" takes on a multitude of meanings. This ability to disconnect the *intended* interpretation from the *available* interpretations is the bedrock of model theory. Model theory is the study of *all* interpretations of a theory, not just the ones that the original author intended.

Of course, model theory isn't really about finding surprising new interpretations — it's much more general than that. It's about exploring the breadth of interpretations that a given theory makes available. It's about discerning properties that hold in all possible interpretations of a theory. It's about discovering how well (or poorly) a given theory constrains its interpretations. It's a toolset used to discuss interpretations in general.

At its core, model theory is the study of what a mathematical theory actually says, when you strip the intent from the symbols.

Iron walls

Before you can do model theory, you have to erect iron walls between four different concepts.

1. Logics
2. Languages
3. Theories
4. Models

Logics

A logic is a formal system for building and manipulating sentences. Traditionally, this logic defines a number of symbols (`( ) ∧ ¬ ∀ ∃ ≡ ν '`, for example) and rules for building sentences from those symbols.

Note that *you cannot generate sentences from a logic alone*. Rather, you *use* a logic to generate sentences *from* a language.

Also, remember that the rules of a logic are *syntactic*, such as "if `φ` is a sentence then `(¬φ)` is a sentence".

Finally, remember that logics are just rules for generating sentences. A logic is perfectly happy to generate sentences shaped like `x∧(¬x)`, in spite of all your protests about contradictions.

Languages

A language is a collection of symbols. *From* those symbols, *using* a logic, you can start generating sentences.

For example, in the propositional logic, using the language `{x, y}`, the string `hello` is surely not a sentence (for it fails to use the appropriate symbols). Nor is the string `¬xy` a sentence: it fails to follow the rules of the logic. `((¬x)∧y)` is a sentence, for it uses the appropriate symbols and follows the given rules.

Many results in model theory are achieved by holding the logic fixed and varying the language, so it's essential that these concepts be distinct in your mind.

Theories

A theory is a collection of sentences written in one language. For example, in the language `{≤}` under first-order logic, we can discuss the theory

1. `(∀x)(x≤x)`
2. `(∀xy)(x≤y)∧(y≤x)→(y≡x)`
3. `(∀xyz)(x≤y)∧(y≤z)→(x≤z)`

which is the theory of *order*. (The axioms above are reflexivity, antisymmetry, and transitivity).

Remember that a theory is just a set of sentences drawn from all available sentences. These sentences aren't particularly special unless you make them special. Sentences like `(∃x)¬(x≤x)` are fine sentences built from the language `{≤}`, even though they directly contradict the theory. Theories don't affect the sentences of a language — they're just a grab-bag of some sentences that seemed interesting to someone.

Models

A model is an *interpretation* of the sentences generated by a language. A model is a structure which assigns a truth value to each sentence generated by some language under some logic.

(More specifically, it's a structure that assigns binary values to sentences in such a way that we're justified in the name "truth value": for example, we require that a model says φ is true if and only if it says that ¬φ is false, and so on.)

Only once we start interpreting sentences is it meaningful to talk about valid or refutable sentences. Once you have a model of {≤} that happens to say that the axioms 1, 2, and 3 above are true, *then* you can start talking about how the theory of order rules out the sentence (∃x)¬(x≤x) — because there is no model of the theory of order which is also a model of this sentence.

(You can actually talk about how (∃x)¬(x≤x) is inconsistent with the theory of order without appealing to model theory, but I find it helpful to treat everything as raw symbols until interpreted by a model.)

To give a concrete example, in *first order logic,* using the *language* {S, +, *, 0}, the *theory* of arithmetic is the theory laid out by the [Peano axioms](http://en.wikipedia.org/wiki/Peano_axioms#First-order_theory_of_arithmetic). The actual natural numbers zero, one, two, ... are a model of this theory (where zero is the interpretation of 0, one is the interpretation of S0, etc.).

Also, it's worth noting that *any* object that interprets sentences and follows the rules of the logic qualifies as a model. There are often many non-isomorphic objects that interpret the same sentences in the same way. For example, rational numbers and real numbers are models of group theory that agree on every sentence in the language of groups, despite being different models.

Distinctions between these four points is something that seems obvious to me in hindsight, but I explicitly remember expending cognitive effort to separate these concepts mentally, so there you go. Make sure these distinctions are wrought in iron before attempting model theory.

The Right to use a name

There's something about math education in general that has troubled me for quite some time, and which I'm finally able to articulate. It's quite possible that this is a personal nit, since nobody else seems to care — but I'll share it anyway.

Many math textbooks treat properties that *justify a name* of a thing as statements about the thing *after* naming it.

This is a little abstract, so I'll make a silly example. Imagine someone is trying to show that, in category theory, composition of arrows is associative. They shouldn't appeal to visual intuition or any diagrams of arrows.

The concept that following arrows is an associative operation is so ingrained in the concept of "arrow" that it's difficult to describe the property in English without sounding dumb.

> If you move from A to B, then move B-to-D-through-C in one step, and if I follow the same paths but move A-to-C-through-B in one step and then from C to D, then we will end up at the same place.

This property of arrows is so stupidly obvious that the statement is frustrating. Further, it hides the following fact:

Associative composition between thingies is something we must have before we're justified in calling the thingies "*Arrows*".

Associative composition is what *allows* you to use the name "arrow" and draw visual diagrams. You can't appeal to my intuition about "arrows" to show that composition is associative. It's the other way around! Only *after* you show that your thingies have associative composition are you allowed to label them as "arrows".

As another example, the axioms of order (above) are what *allow* us to use the $\leq$ symbol, which appeals to our intuitive idea of order. Really, it's more honest to say "We have a binary relation R, satisfying

1. $(\forall x)R(xx)$
2. $(\forall xy)R(xy)\wedge R(yx)\rightarrow(y\equiv x)$
3. $(\forall xyz)R(xy)\wedge R(yz)\rightarrow R(xz)$

which *justifies* our use of the $\leq$ symbol for R."

I imagine this is not a problem for experienced mathematicians, for whom it goes without saying that you must formally specify (or disregard) all intuitive baggage that comes attached to the names. However, I remember distinctly a number of times when I gnashed my teeth with boredom as teachers made obvious statements (*of course* $\leq$ is reflexive, why do we even need to say this?), simply because I didn't understand this idea.

I mention this because the first few sections of the *Model Theory* textbook make statements that seem quite obvious. It's easy to grind your teeth and say "duh, hurry up". It's a little harder to understand exactly why such things must be said. In that light, I think this is a good piece of advice for learning mathematics in general:

If you find yourself wondering why a statement must be said, check whether the statement is justifying any names.

Binding meaning

The early parts of *Model Theory* will go down much easier if you realize that they're binding logical symbols to the appropriate meaning (and thus justifying the name "model").

For example, when we state "M models $\varphi\wedge\psi$ if and only if it models φ and it models ψ", it's easy to say "well duh". It's a little harder to understand that *this is the mechanism by which the symbol* $\wedge$ *is bound to the interpretation* "*and*".

Also, note that the ability to distinguish between "the symbol $+$ in the language L" from "the addition function as interpreted by the model M" is absolutely crucial.

Totality

Something that kept on biting me was this: *Models of first-order logic are* "*total*". They have something to say about *every* sentence in a language. Even where a *theory* is incomplete, any individual *model* is "complete". A model of first-order logic interprets function symbols by total functions and relations by set-theoretic relations. The relationship $\models$ is total: for every sentence, either $M\models\varphi$ or $M\models\neg\varphi$.

This is a point where my intuitive notion of "models as interpretations" departed from the actual mathematical objects under consideration — functions are firmly partial-by-default in my mind's eye.

It's important to hold firm the distinction between "model" and "theory" here. Remember that the number *theory* is incomplete, while the standard *model* of number theory is the one that picks "true" for all Gödel sentences, has no infinite numbers, etc. (The difficulties in pinpointing such a model is exactly what the incompleteness theorem is all about.)

Be aware that the mathematical definition of a model may not match your intuitive idea of "a structure which interprets a theory", *especially* if you're coming from computer science (or other constructive fields).

None of this is particularly novel. Rather, this is a collection of distinctions and clarifications that would have made my life a bit easier when beginning the textbook.

In my case, I didn't have any of these concepts wrong, per se — rather, I had them fuzzy. The above distinctions were not yet fleshed out in my mind. This post provides a context for model theory; a taste of the type of thinking you must be ready to think.

I was originally going to use this as context for what I've learned in model theory so far, but this post took longer than expected. I'll follow up tomorrow.

How habits work and how you may control them

Some highlights from The Power of Habit: Why We Do What We Do in Life And Business by Charles Duhigg, a book which seems like an invaluable resource for pretty much everyone who wants to improve their lives. The below summarizes the first three chapters of the book, as well as the appendix, for I found those to be the most valuable and generally applicable parts. These chapters discuss individual habits, while the rest of the book discusses the habits of companies and individuals. The later chapters also contain plenty of interesting content (some excerpts: [1 2 3]), and help explain the nature of e.g. some institutional failures.

(See also two previous LW discussions on an online article by the author of the book.)

Chapter One: **The Habit Loop - How Habits Work**

When a rat first navigates a foreign environment, such as a maze, its brain is full of activity as it works to process the new environment and to learn all the environmental cues. As the environment becomes more familiar, the rat's brain becomes less and less active, until even brain structures related to memory quiet down a week later. Navigating the maze no longer requires higher processing: it has become an automatic habit.

The process of converting a complicated sequence of actions into an automatic routine is known as "chunking", and human brains carry out a similar process. They vary in complexity, from putting toothpaste on your toothbrush before putting it in your mouth, to getting dressed or preparing breakfast, to very complicated processes such as backing one's car out of the driveway. All of these actions initially required considerable effort to learn, but eventually they became so automatic as to be carried out without conscious attention. As soon as we identify the right cue, such as pulling out the car keys, our brain activates the stored habit and lets our conscious minds focus on something else. In order to conserve effort, the brain will attempt to turn almost any routine into a habit.

However, it can be dangerous to deactivate our brains at the wrong time, for there may be something unanticipated in the environment that will turn a previously-safe routine into something life-threatening. To help avoid such situations, our brains evaluate prospective habits using a three-stage *habit loop*:

> From behind a partition, for instance, it’s difficult for a rat to know if it’s inside a familiar maze or an unfamiliar cupboard with a cat lurking outside. To deal with this uncertainty, the brain spends a lot of effort at the beginning of a habit looking for something— a cue— that offers a hint as to which pattern to use. From behind a partition, if a rat hears a click, it knows to use the maze habit. If it hears a meow, it chooses a different pattern. And at the end of the activity, when the reward appears, the brain shakes itself awake and makes sure everything unfolded as expected.
>
> This process within our brains is a three-step loop. First, there is a *cue*, a trigger that tells your brain to go into automatic mode and which habit to use. Then there is the *routine*, which can be physical or mental or emotional. Finally, there is a

reward, which helps your brain figure out if this particular loop is worth remembering for the future.

Over time, this loop— cue, routine, reward; cue, routine, reward— becomes more and more automatic. The cue and reward become intertwined until a powerful sense of anticipation and craving emerges. Eventually, whether in a chilly MIT laboratory or your driveway, a habit is born.

Unused habits disappear very slowly, if at all. If a rat is trained to find cheese in a particular section of the maze, and the cheese is then moved to a different location, it will obtain a new habit. But once the cheese is moved back to its original location, the old habit re-emerges, almost as if it had been active for the whole time. This is part of the reason why it is so hard to start exercising regularly, or to change one's diet: the habit of relaxing in front of the TV, or snacking on a meal, will still be activated by the old cues and engage the behavioral pattern. On the other hand, if one does manage to establish a habit of ignoring the snacks or going out for a jog, it will eventually become as automatic as any other habit.

Habits are crucial for our ability to function. People with damage to the basal ganglia, the parts of the brain responsible for habitual behavior, often become mentally paralyzed. Even basic activities, such as opening a door or choosing what to eat, become difficult to perform, and they may need to pause to wonder whether they should tie their left or right foot first, or whether to brush their teeth before or after taking a shower.

In one set of experiments, for example, researchers affiliated with the National Institute on Alcohol Abuse and Alcoholism trained mice to press levers in response to certain cues until the behavior became a habit. The mice were always rewarded with food. Then, the scientists poisoned the food so that it made the animals violently ill, or electrified the floor, so that when the mice walked toward their reward they received a shock. The mice knew the food and cage were dangerous — when they were offered the poisoned pellets in a bowl or saw the electrified floor panels, they stayed away. When they saw their old cues, however, they unthinkingly pressed the lever and ate the food, or they walked across the floor, even as they vomited or jumped from the electricity. The habit was so ingrained the mice couldn’t stop themselves.

It’s not hard to find an analog in the human world. Consider fast food, for instance. It makes sense— when the kids are starving and you’re driving home after a long day— to stop, just this once, at McDonald’s or Burger King. The meals are inexpensive. It tastes so good. After all, one dose of processed meat, salty fries, and sugary soda poses a relatively small health risk, right? It’s not like you do it all the time.

But habits emerge without our permission. Studies indicate that families usually don’t intend to eat fast food on a regular basis. What happens is that a once a month pattern slowly becomes once a week, and then twice a week— as the cues and rewards create a habit— until the kids are consuming an unhealthy amount of hamburgers and fries. When researchers at the University of North Texas and Yale tried to understand why families gradually increased their fast food consumption, they found a series of cues and rewards that most customers never knew were influencing their behaviors. They discovered the habit loop.

Every McDonald’s, for instance, looks the same— the company deliberately tries

to standardize stores’ architecture and what employees say to customers, so everything is a consistent cue to trigger eating routines. The foods at some chains are specifically engineered to deliver immediate rewards— the fries, for instance, are designed to begin disintegrating the moment they hit your tongue, in order to deliver a hit of salt and grease as fast as possible, causing your pleasure centers to light up and your brain to lock in the pattern. All the better for tightening the habit loop.

However, even these habits are delicate. When a fast food restaurant closes down, the families that previously ate there will often start having dinner at home, rather than seek out an alternative location. Even small shifts can end the pattern. But since we often don’t recognize these habit loops as they grow, we are blind to our ability to control them. By learning to observe the cues and rewards, though, we can change the routines.

Chapter Two: The Craving Brain - How to Create New Habits

A basic rule of marketing, based on the habit loop, is to attempt to identify a simple obvious cue, and then offer a clear reward from one's product. An early success was in the marketing of Pepsodent, where the marketer instructed people to run their tongue across their teeth and notice the existence of a "film" on the teeth. He then argued that by using his toothpaste, people could get rid of the film and obtain beautiful, clean teeth. (In reality, the "film" is a harmless membrane that builds up on teeth regardless of how often one eats or brushes their teeth.)

However, other toothpaste companies had tried similar marketing tactics before, without much success. Another part of Pepsodent's success was that it happened to contain citric acid, as well as other chemicals that act as mild irritants. Their effect is to create a cool, tingling sensation on the tongue and gums of people. This acted as the real reward for the habit - although the sensation itself only happened to occur by coincidence, people came to associate it with having brushed their teeth, and of having a clean mouth. It was when people began craving this reward that tooth brushing really became a habit. When other toothpaste companies realized what was going on, they all proceeded to add similar irritants to their products.

“Consumers need some kind of signal that a product is working,” Tracy Sinclair, who was a brand manager for Oral-B and Crest Kids Toothpaste, told me. “We can make toothpaste taste like anything— blueberries, green tea— and as long as it has a cool tingle, people feel like their mouth is clean. The tingling doesn’t make the toothpaste work any better. It just convinces people it’s doing the job.”

When a habit becomes sufficiently established in the brain, the cue no longer just activates the routine - it also makes us crave the reward that is associated with completing the routine. If the cue is present, but we can't engage in the routine or try to prevent ourselves from doing so, the craving will increase in strength until it becomes almost overpowering. Various cues - the sight of a pack of cigarettes, the smell of food, a computer or smartphone chiming to signify the arrival of a new message - can activate the anticipatory mechanism, and the craving to take a smoke, eat a bite, or check one's messages.

Scientists have studied the brains of alcoholics, smokers, and over-eaters and have measured how their neurology— the structures of their brains and the flow of neurochemicals inside their skulls— changes as their cravings became ingrained. Particularly strong habits, wrote two researchers at the University of Michigan,

> produce addiction-like reactions so that “wanting evolves into obsessive craving” that can force our brains into autopilot, “even in the face of strong disincentives, including loss of reputation, job, home, and family".

The same mechanisms can also be used to encourage good or healthy habits. One chooses a cue, such as going to the gym as soon as one wakes up, and a reward, such as smoothie after each workout. Then one thinks about the smoothie, or the endorphin rush that follows during the exercise. As one allows oneself to anticipate the reward, a craving will begin to ensue, which will make it easier to get oneself to the gym every day. (See also PJ Eby on this.)

> Cravings are what drive habits. And figuring out how to spark a craving makes creating a new habit easier. It’s as true now as it was almost a century ago. Every night, millions of people scrub their teeth in order to get a tingling feeling; every morning, millions put on their jogging shoes to capture an endorphin rush they’ve learned to crave.

Chapter Three: The Golden Rule of Habit Change - Why Transformation Occurs.

The Golden Rule of Habit Change is that one cannot extinguish a bad habit, only change it. One keeps the old cue and the old reward, but changes the routine. Almost any behavior can be changed if the cue and reward stay the same.

For example, alcoholics rarely crave the actual physical state of intoxication itself. Rather, people drink in order to obtain escape, relaxation, companionship, blunting of anxieties, or an opportunity for emotional release. Organizations such as Alcoholics Anonymous[1] build a system of "sponsors" and group meetings, allowing a person in need of relief to talk with their sponsor or attend a group meeting. The cue, a need for relief, stays the same, as does the reward: getting relief. What changes is the behavior: instead of drinking, one obtains their relief by talking to others.

Habit reversal therapy is the formal version of this technique. In one example, Mandy, a 24-year-old graduate student had a compulsive need to bite her nails. The therapist asked Mandy to describe what she felt right before bringing her hand up to her mouth to bite her nails: Mandy described experiencing a feeling of tension. This was the cue for the habit. After some discussion, they established that Mandy bit her fingers when she was bored, and after she had worked through all of her nails, she felt a brief sense of completion. The physical stimulation acted as the reward.

> At the end of their first session, the therapist sent Mandy home with an assignment: Carry around an index card, and each time you feel the cue— a tension in your fingertips— make a check mark on the card. She came back a week later with twenty-eight checks. She was, by that point, acutely aware of the sensations that preceded her habit. She knew how many times it occurred during class or while watching television.
>
> Then the therapist taught Mandy what is known as a “competing response.” Whenever she felt that tension in her fingertips, he told her, she should immediately put her hands in her pockets or under her legs, or grip a pencil or something else that made it impossible to put her fingers in her mouth. Then Mandy was to search for something that would provide a quick physical stimulation— such as rubbing her arm or rapping her knuckles on a desk— anything that would produce a physical response.

The cues and rewards stayed the same. Only the routine changed.

They practiced in the therapist's office for about thirty minutes and Mandy was sent home with a new assignment: Continue with the index card, but make a check when you feel the tension in your fingertips and a hash mark when you successfully override the habit.

A week later, Mandy had bitten her nails only three times and had used the competing response seven times. She rewarded herself with a manicure, but kept using the note cards. After a month, the nail-biting habit was gone. The competing routines had become automatic. One habit had replaced another.

[…]

Say you want to stop snacking at work. Is the reward you're seeking to satisfy your hunger? Or is it to interrupt boredom? If you snack for a brief release, you can easily find another routine— such as taking a quick walk, or giving yourself three minutes on the Internet— that provides the same interruption without adding to your waistline.

If you want to stop smoking, ask yourself, do you do it because you love nicotine, or because it provides a burst of stimulation, a structure to your day, a way to socialize? If you smoke because you need stimulation, studies indicate that some caffeine in the afternoon can increase the odds you'll quit. More than three dozen studies of former smokers have found that identifying the cues and rewards they associate with cigarettes, and then choosing new routines that provide similar payoffs— a piece of Nicorette, a quick series of push-ups, or simply taking a few minutes to stretch and relax— makes it more likely they will quit.

For some habits, though, this is not enough. The alcoholics who replace their old behaviors with new ones may manage to stop drinking for a long while, until they run into some particularly stressful event in their lives. At this point, the stress becomes too much for many, who start drinking again. Not everyone does, however, and the difference seems to be in whether people are capable of genuinely believing that things will become better.

However, those alcoholics who believed, like John in Brooklyn, that some higher power had entered their lives were more likely to make it through the stressful periods with their sobriety intact.

It wasn't God that mattered, the researchers figured out. It was belief itself that made a difference. Once people learned how to believe in something, that skill started spilling over to other parts of their lives, until they started believing they could change. Belief was the ingredient that made a reworked habit loop into a permanent behavior.

"I wouldn't have said this a year ago— that's how fast our understanding is changing," said Tonigan, the University of New Mexico researcher, "but belief seems critical. You don't have to believe in God, but you do need the capacity to believe that things will get better.

"Even if you give people better habits, it doesn't repair why they started drinking in the first place. Eventually they'll have a bad day, and no new routine is going to

make everything seem okay. What can make a difference is *believing* that they can cope with that stress without alcohol."

Appendix: A Reader's Guide to Using These Ideas

There isn't a single formula for changing habits, but rather thousands. Different people are driven by different cravings, and different habits require different approaches: stopping overeating is different from giving up cigarettes, which is different from how one communicates with their spouse. That said, the author attempts to provide a general framework for changing habits. It consists of four steps: Identify the routine, experiment with rewards, isolate the cue, have a plan.

The routine involved in the habit is usually the most obvious aspect. For example, maybe somebody always gets up from their desk at afternoon, walks to a cafeteria, buys a cookie, and eats it while chatting with friends. What exactly is the reward here? It could be the cookie itself, the change of scenery, the temporary distraction, the opportunity to socialize with colleagues, or the burst of energy that comes from the blast of sugar.

To identify the answer, one needs to experiment with rewards. On one day, instead of going out to a cafeteria, they might instead take a walk around the block. Another day, they might go to the cafeteria and buy an apple or chocolate bar and return to their desk without talking to anyone. On yet another day, they might walk to someone's desk to gossip for a few minutes and then return to work. When they do return to their desk, they should take a moment to quickly write down their thoughts or feelings - even just in the form of three random words in their head, like "relaxed", "saw flowers", "not hungry" - and then set a fifteen-minute alarm. If, after fifteen minutes, they still feel the craving, they know that whatever it was that they just did, it didn't give the desired reward. On the other hand, if they replaced the cafeteria visit by going to chat with a friend and the cafeteria craving vanished, then they've identified the reward as being a desire for temporary distraction and socialization.

Then there is the task of identifying the cue. Experiments have shown that almost all habitual cues fall into one of five categories:

1. Location
2. Time
3. Emotional state
4. Other people
5. Immediately preceding action

So when one notices themselves engaging in a habit, they can write down the state of each of these variables. For example, here's one of the notes that the author made while trying to diagnose his own snacking habit:

> Where are you? (sitting at my desk)
>
> What time is it? (3:36 P.M.)
>
> What's your emotional state? (bored)
>
> Who else is around? (no one)
>
> What action preceded the urge? (answered an e-mail)

After making such notes for three days, the pattern became clear: he got an urge to snack sometime between 3:00 and 4:00. The reward was temporary distraction, the kind that comes from gossiping with a friend.

Now he needed to have a plan for overriding the old habit with a new one, while maintaining the old cue and reward. So he wrote down the following:

> At 3:30, every day, I will walk to a friend's desk and talk for 10 minutes.

To make sure I remembered to do this, I set the alarm on my watch for 3: 30.

It didn’t work immediately. There were some days I was too busy and ignored the alarm, and then fell off the wagon. Other times it seemed like too much work to find a friend willing to chat— it was easier to get a cookie, and so I gave in to the urge. But on those days that I abided by my plan— when my alarm went off, I forced myself to walk to a friend’s desk and chat for ten minutes— I found that I ended the workday feeling better. I hadn’t gone to the cafeteria, I hadn’t eat a cookie, and I felt fine. Eventually, it got be automatic: when the alarm rang, I found a friend and ended the day feeling a small, but real, sense of accomplishment. After a few weeks, I hardly thought about the routine anymore. And when I couldn’t find anyone to chat with, I went to the cafeteria and bought tea and drank it with friends.

That all happened about six months ago. I don’t have my watch anymore— I lost it at some point. But at about 3:30 every day, I absentmindedly stand up, look around the newsroom for someone to talk to, spend ten minutes gossiping about the news, and then go back to my desk. It occurs almost without me thinking about it. It has become a habit.

Footnotes

[1]: How effective is the AA? The book admits that the effectiveness is hard to evaluate, but notes that *An estimated 2.1 million people seek help from AA each year, and as many as 10 million alcoholics may have achieved sobriety through the group. AA doesn’t work for everyone— success rates are difficult to measure, because of participants’ anonymity— but millions credit the program with saving their lives.* It also comments that although scientists have been critical of the AA's unscientific methodology in the past, increasing numbers of researchers have recently become interested in the organization as its methodology fits other findings about habit change.

Bayesianism for Humans

Recently, I completed my first systematic read-through of the sequences. One of the biggest effects this had on me was considerably warming my attitude towards Bayesianism. Not long ago, if you'd asked me my opinion of Bayesianism, I'd probably have said something like, "Bayes' theorem is all well and good when you know what numbers to plug in, but all too often you don't."

Now I realize that that objection is based on a misunderstanding of Bayesianism, or at least Bayesianism-as-advocated-by-Eliezer-Yudkowsky. "When (Not) To Use Probabilities" is all about this issue, but a cleaner expression of Eliezer's true view may be this quote from "Beautiful Probability":

> No, you can't always do the exact Bayesian calculation for a problem. Sometimes you must seek an approximation; often, indeed. This doesn't mean that probability theory has ceased to apply, any more than your inability to calculate the aerodynamics of a 747 on an atom-by-atom basis implies that the 747 is not made out of atoms. Whatever approximation you use, it works to the extent that it approximates the ideal Bayesian calculation - and fails to the extent that it departs.

The practical upshot of seeing Bayesianism as an ideal to be approximated, I think, is this: **you should avoid engaging in any reasoning that's demonstrably nonsensical in Bayesian terms.** Furthermore, **Bayesian reasoning can be fruitfully mined for heuristics that are useful in the real world.** That's an idea that actually has real-world applications for human beings, hence the title of this post, "Bayesianism for Humans."

Here's my attempt to make an initial list of more directly applicable corollaries to Bayesianism. Many of these corollaries are non-obvious, yet eminently sensible once you think about them, which I think makes for a far better argument for Bayesianism than Dutch Book-type arguments with little real-world relevance. Most (but not all) of the links are to posts within the sequences, which hopefully will allow this post to double as a decent introductory guide to the parts of the sequences that explain Bayesianism.

- Watch out for base rate neglect. It's why even experts screw up in one of the standard problems used to explain Bayes' Theorem. Even when you don't know what the base rate is, there are times when you ought to expect it to be low, particularly if you're trying to detect a rare phenomenon like a new disease or IDing terrorists.
- Absence of Evidence is Evidence of Absence. If observing E would increase the probability of H, observing not-E should decrease the probability of H. E and not-E sure as hell shouldn't *both* increase the probability of H.
- Relatedly, there's Conservation of Expected Evidence: roughly, if you think more evidence would probably increase your confidence in a belief, you should think there's small chance it would cause a larger change in the opposite direction.
- Conservation of expected evidence means that a rational person can't seek to confirm their beliefs, only to test them. If your expectation of how a test will affect your belief violates conservation of expected evidence, you should update your beliefs *now* based on how you expect the test to turn out.

- Also related closely related to the above, when it comes to gathering evidence, "If you know your destination, you are already there." Evidence gathered through a biased method designed to turn out one way is worthless.
- On the other hand, you can't dismiss a hypothesis due to a lack of a *particular* piece of evidence that you wouldn't expect to have even if the hypothesis were true.
- Even when an argument on one side is overcome by a stronger argument on the other side, you still need to take the first argument into account when assigning confidence to your belief, lest you gradually dismiss each piece of evidence on the other side because no piece is (individually) as strong as the one piece of evidence on your side.
- Burdensome Details: Every detail added to a claim makes it less probable.
- Reversed Stupidity Is Not Intelligence: If you would expect to see flying saucer cults regardless of whether or not extraterrestrials were visiting us, flying saucer cults are not evidence against extraterrestrials.
- Don't get caught up in arguing about definitions when you should be looking at what's actually indicative of what. (I take it that that's the Bayesian take-away from the sequence on words, though Eliezer doesn't quite put it that way.)
- Rationality and the rules of Science are not the same thing. The latter are social rules designed to make science work in spite of the irrationality of its practitioners. They're not the same as the rules of rationality an ideal reasoner would follow.
- An example of the science vs. rationality issue: to an ideal reasoner, successful retrospective predictions are as valuable as prospective predictions. (To us non-ideal reasoners, prospective predictions are can be extra valuable as protection against fooling ourselves, but we still shouldn't discount retrospective predictions entirely.)
- This is also reason to be careful about dismissing evo psych claims as just-so stories.
- Another example: contrary to old-fashioned statistical procedure, a researcher's state of mind shouldn't affect the significance of their results.
- Last example is from a post of my own: Expert opinion should be discounted when the expert's opinions could be predicted solely from information not relevant to the truth of the claims. But when the state of expert opinion surprises you, beware discounting their opinions just because you can think of *some* explanation for why they'd be wrong.

[Link] Low-Hanging Poop

Related: Son of Low Hanging Fruit

Another post on finding low hanging fruit from Gregory Cochran's and Henry Harpending's blog West Hunter.

> *Clostridium difficile* causes a potentially serious kind of diarrhea triggered by antibiotic treatments. When the normal bacterial flora of the colon are hammered by a broad-spectrum antibiotic, *C. difficile* often takes over and causes real trouble. Mild cases are treated by discontinuing antibiotic therapy, which often works: if not, the doctors try oral metronidazole (Flagyl), then vancomycin , then intravenous metronidazole. This doesn't always work, and *C. difficile* infections kill about 14,000 people a year in the US.
>
> One recent trial shows that fecal bacteriotherapy, more commonly called a stool transplant, works like gangbusters, curing ~94% of patients. The trial was halted because the treatment worked so well that refusing to poopify the control group was clearly unethical. I read about this, but thought I'd heard about such stool transplants some time ago. I had. It was mentioned in *The Making of a Surgeon*, by William Nolen, published in 1970. Some crazy intern – let us call him Hogan – tried a stool transplant on a woman with a *C. difficile infection.* He mixed some normal stool with chocolate milk and fed it to the lady. It made his boss so mad that he was dropped from the program at the end of the year. It also worked. It was inspired by a article in *Annals of Surgery,* so this certainly wasn't the first try. According to Wiki, there are more than 150 published reports on stool transplant, going back to 1958.
>
> So what took so damn long? Here we have a simple, cheap, highly effective treatment for *C. difficile* infection that has only become officially valid this year. Judging from the *H. pylori* story, it may still take years before it is in general use.
>
> Obviously, sheer disgust made it hard for doctors to embrace this treatment. There's a lesson here: in the search for low-hanging fruit, reconsider approaches that are embarrassing, or offensive, or downright disgusting.
>
> **Investigate methods were abandoned because people hated them, rather because of solid evidence showing that they didn't work.**
>
> Along those lines, no modern educational reformer utters a single syllable about corporal punishment: doesn't that make you suspect it's effective? I mean, why we aren't we caning kids anymore? The Egyptians said that a boy's ears are in his back: if you do not beat him he will not listen. Maybe they knew a thing or three.
>
> Sometimes, we hate the idea's *authors*: the more we hate them, the more likely we are to miss out on their correct insights. Even famous assholes had to be competent in *some* areas, or they wouldn't have been able to cause serious trouble.

How to Learn from Experts

The key difference between experts and beginners is the quality of their abstractions. Masters of a field mentally organize information in a way that's relevant to the tasks at hand. Amateurs may know as many facts and details as experts but group them in haphazard or irrelevant ways.

For example, experienced Bridge players group cards by suit, then number. They place the most importance on the face cards and work down. Bridge amateurs group solely by number and place equal importance on all numbers. Professional firemen group fires by how the fire was started and how fast it's spreading-features they use to contain the fire. Novices group fires by brightness and color. Both have the same information, but the firemen hone in on the useful details faster.[1]

Learn abstractions from masters. If you ask a Software Architect which database technology you should use, circumstances will eventually change and you'll need to ask them again and pay them again. But if you ask the Architect to teach you how to choose a database then you can adapt to changing circumstances. **Ideally you should emerge with a clear set of rules**-something like a flow-chart for that decision. A good example is this article on whether you should use hadoop. Clear criteria let you make a high-quality decision by focusing on the relevant details.

After talking to the expert you can write up the flow-chart or criteria and send it to them to get their opinion. This ensures you understood what the expert was trying to say, and lets you get additional details they might add. Most importantly it gives them something valuable to share with people seeking similar advice, so you're able to add value to their lives as a thank you for their advice.

Caveats to this method:

- In some domains there are details only professionals know. Academic research has a secret paper-passing network with ideas known to top researchers 1-2 years before they're published. So you need to be in constant contact with these experts and hear the details from them. However, this typically only matters if you're aiming to become a top-class expert yourself.
- Experts aren't always conscious of the abstractions they use. They'll say one thing and do another. So you should ask them to guide you through a specific situation and ask them several questions about how their decision would change if some conditions are different.
- You may not have a specific question you want answered-you might want to find "unknown unknowns." In that case ask the expert for stories-things they did that made a big difference. Then analyze those situations to figure out what criteria they used.

1: *The Cambridge Handbook of Expertise and Expert Performance*

What Can We Learn About Human Psychology from Christian Apologetics?

A couple months ago I set up a Skype meeting Robin Hanson to chat about the book he's working on. But the first thing he wanted to talk wasn't directly related to the book. He'd read some of my work critiquing Christian apologetics, and said something to the effect of even though people who spend a lot of time arguing about religion are extreme cases, maybe they somehow shed light on the psychology of ordinary people. I didn't have a good response at the time; I had taken a shot at discussing the sociology of apologetics in my first book, but I was never terribly satisfied with that chapter and hadn't thought about the subject much since writing it.

Since then, I've thought about it more, and now have a better answer for Robin. The take-away is that to understand Christian apologetics, you need to see it as a giant exercise is violating Eliezer's advice in the Against Rationalization subsequence, particularly The Bottom Line. What's particularly noteworthy is the enormous amount of effort many Christians put into doing so, rather than just shrugging their shoulders and saying "I believe on faith." (Note: everything I say here is probably applicable to some degree to other forms of apologetics, but I'll focus on Christian apologetics and in particular *Protestant* apologetics because it's what I'm most familiar with.)

And I need to emphasize from the start that we are talking about a lot of Christians here. Big name professional apologists are rare, but then so by definition are "big names" in any field. *Consumers* of apologetics are not so rare: countless evangelicals have read C. S. Lewis' apologetic work *Mere Christianity* and it's number 3 on *Christianity Today's* list of "The Top 50 Books That Have Shaped Evangelicals." Immediately following *Mere Christianity* on the list is another apologetic work, Francis Schaeffer's *The God Who Is There.* Josh McDowell's *The Evidence That Demands A Verdict* is number 13. McDowell's *More Than A Carpenter* has reportedly sold 15 million copies, while Lee Strobel's *Case For...* books have reportedly sold 10 million copies all together.

I think apologists are best-seen as a highly specialized kind of religious professional, in some ways analogous to priests and ministers. Indeed there's overlap: many prominent apologists have had less well-known careers as pastors, while many evangelical pastors brush up on their apologetic arguments to share them with their congregations.

The second thing you need to understand, if you want to make sense of apologetics, is that apologists are in the business of pretending the purpose of apologetics is something other than what it actually is. This is not something you will learn *even from reading many atheist critiques of apologetics,* because many critics are willing to politely play along with the pretense that the purpose of apologetics is to open minded-skeptics and debates between skeptics and believers are serious intellectual engagements.

Such politeness may actually be smart tactics, if you are addressing believers and your goal is to persuade them, but that's not what I'll be doing here. Instead, I'll be

addressing the mostly-atheist readership of LessWrong, and my goal will be to see what we can learn from apologetics about human psychology in general.

One of the best discussions I've read of the false pretenses of apologetics is a relatively brief section in Robert J. Miller's commentary on a debate between evangelical apologist William Lane Craig and liberal Christian scholar John Dominic Crossan (published alongside other commentaries and a transcript of the debate as *Will the Real Jesus Please Stand Up?*). Miller writes:

> Why is it that few, if any, outsiders will be persuaded by Craig's apology? From the way he presents it, we get the impression that he thinks nobody who is informed, rational, and sincere could disagree with it...
>
> I used to think this way myself when I was a fervent believer in the power of apologetics. I was a philosophy major at a Catholic college. I was utterly convinced not only that Christianity was the one true religion that God intended for all humanity, but also that the Catholic Church was the one true church that Christ intended for all Christians. From my study of Thomas Aquinas and modern Christian apologetics, I clearly saw that the central truths of Christianity (and of Catholicism) could be grasped by reason if only one was sincerely seeking God's truth, was humble enough to accept it, and took the time to inform oneself and follow the arguments.
>
> All of this made perfect sense to me, and none of my teachers or fellow students (all of whom were Catholics) gave me any reason to question it. I tried out various apologetic arguments on my like-minded friends, who found them quite convincing. Occasionally they suggested improvements in my arguments, but none of us doubted the effectiveness of apologetics. The only real puzzle in my mind was this: since the truths of Christianity and Catholicism are so evident, why are they not more universally recognized? I concluded that those outside my religion or my church just did not know or did not understand these apologetic arguments, or that they were not completely sincere about seeking the truth...
>
> This mind-set held together until I went to graduate school at secular universities and got to know people who had different religions. For the first time in my life, I got to know people who took other religions as seriously as I took mine. I knew these people were well educated and highly rational, and I could tell from our conversations that they were sincere. A few were people of great goodness and spiritual depth. Yet none of them was persuaded by my apologetics.

This means that if the purpose of apologetics is taken at face-value, "apologies are almost always abject failures." However, he writes:

> The is another, more promising way to evaluate the apologetic genre. We can determine its audience, not by whom it seems to be aimed at, but by who actually reads it. And we can determine its purpose, not by what the author seems to intended, but by how it actually functions. if we proceed like this, we reach two important findings: (1) the audience for an apology is insiders; (2) its function is to support what the audience already believes.
>
> This is nothing new to apologists, who know full well that their audiences are insiders. (Why else would Craig speak at Moody Memorial Church or write for Baker Book House?) So why do apologists write as if they were addressing outsiders? They do that, not because they are mistaken about their audience, but because that is the convention of the apologetic genre. An apt comparison is the

> genre of the open letter. An open letter may begin, "To the President of the United States," but both author and readers understand that the real audience is the general public. Readers don't think they are reading the president's mail... Authors of fables write about talking animals because that is how fables go, not because anyone thinks that animals really talk.

While Miller makes good points, he is too kind to treat the pretense of persuading outsiders as a mere genre convention and imply nobody believes it. He certainly seems to have believed his arguments would persuade outsiders when he was a Catholic college student.

Furthermore, both Josh McDowell and Lee Strobel make their self-presentation as former skeptics persuaded by overwhelming evidence a central part of their marketing. Their fans seem to mostly believe the marketing, and would therefore conclude Miller is wrong about the purpose of apologetics. But scratch the surface, and you start to see marketing is all it is. In recent editions of his books, McDowell claims that in college he traveled Europe researching the evidence for Christianity, but I've been unable to find any record of this claim prior to the 1999 edition of *The Evidence that Demands a Verdict* (the first edition was published in 1972).

Lee Strobel's *Case for...* books go even further playing up the "former skeptic" angle. They consist of a series of interviews with Christian apologists, presented in narrative form with Strobel feigning skepticism and objectivity while pitching the apologists softball questions. In my experience, many of Strobel's fans believe their reading an account of Strobel's conversion. More attentive readers will notice Strobel only claims to be "retracing" his conversion. Strobel's earlier book, *Inside the Mind of Unchurched Harry and Mary,* gives the real story: Strobel started going to church because of his wife, found it emotionally moving, and then started reading up on apologetics to assure himself it was all true.

Apologetics is marketed this way because fans of apologetics want to believe it. And in his reply to Miller, Craig tries to keep up that image of apologetics, even while conceding some of Miller's points. Craig says he publishes at with evangelical publishing houses because "it is extraordinarily difficult to interest nonevangelical presses in publishing a defense of the historical resurrection of Jesus." Somehow, Craig doesn't consider that this might be because the audience for such material is composed almost entirely of evangelicals.

Craig concedes that few outsiders will be persuaded by his arguments, but then says there are exceptions to this rule. He has a couple stories of how, after one of his appearances on a college campus, a staff member from a campus Christian org (presumably the one that organized the event) told him he'd made some converts.

He also tells tells a story about meeting an investment banker who says he had "wanted to believe in Jesus," but had trouble buying the resurrection story. So he joined a small group at a local church and spent some time talking to one of the ministers there, who "laid out for him the evidence for Jesus' miraculous resurrection. After reading a book of evangelical responses to the liberal Jesus Seminar, the man says that "I asked Jesus into my life."

But Craig concedes the people in his anecdotes are unusual, so before I say anything about them, let's talk about the majority of apologetics consumers who are already believers. For many, I suspect, apologetics gives them a few extra good feels about their faith, but that's the extent of what it does for them. Miller certainly doesn't make

it sound like his college-age self would have faced a major crisis of faith without apologetics.

For other Christians, however, consuming apologetics is part of a desperate attempt to hold on to their beliefs in the face of doubts. The ranks of the atheist movement are full of ex-Christians who went through an apologetics-reading phase for this reason. My impression, furthermore, is that there are Christians who have succeeded where many current atheists have failed. For example, Christian apologist Mike Licona (who made headlines when he was forced to resign from his position at Southern Evangelical Seminary for his ever-so-slight deviations from the inerrantist party line) credits his mentor in apologetics, Gary Habermas, with saving his faith.

In fact, when I read Eliezer say that, in the Orthodox Judaism of his childhood, "You're allowed to doubt. You're just not allowed to successfully doubt," this struck me as a pretty good expression of an attitude that's common in evangelical Protestant apologetics. They may not take it as far as it's taken in Eliezer's account of Judaism—they don't raise doubts just to have a competition over who can come up with the most complicated explanation—but there's a resigned recognition that doubt is inevitable. So they talk about struggling with doubt, dealing with doubt, overcoming doubt, living with doubt. The message is that doubt can be embraced or at least tolerated, as long as you don't, as Eliezer would put it, doubt successfully.

Apologetics, though, seems to serve another, stranger purpose. Once, in college, I attended an apologetics talk put on by the local Campus Crusade chapter, and after the talk ran into an acquaintance who I got to talking with. He explained friends of his had told him about how Christianity had saved their lives, which made him want to convert, but he wasn't sure he could really believe it, hence going to the talk.

This seems to be part of a pattern with other stories I've heard, like Lee Strobel's story (the relatively unvarnished version from *Inside the Mind...*) and Craig's story of the investment banker: people decide they want to convert for emotional reasons, but some can't believe it at first, so they use apologetics as a tool to get themselves to believe what they've decided they want to believe.

In "The Bottom Line," Eliezer imagines the owner of a box paying a clever arguer to argue that there's a diamond inside. This is, in effect, the role of apologists, to make a living as clever arguers serving people who've decided they want to believe certain religious doctrines are true. As someone who's had rationalist instincts since before I knew anything about rationalism (as an intellectual tradition or movement), part of me is surprised that this would *ever* work. Shouldn't it be obvious to people that they're fooling themselves?

On the other hand, it says something about people's need to feel rational that they would go to the trouble, rather than just satisfying themselves with believing on faith, as many religious believers seem to do. In fact, this need may be more widespread than most people realize. In *Why People Believe Weird Things,* Michael Shermer reports on a study that found that while even most religious believers tend to assume other people believe for non-rational reasons, when you ask religious people about their own reasons for their religious beliefs, they're more likely to cite the argument from design than faith.

(What does all this mean for domains outside religion? I'm not actually sure, though there's some rather obvious connections you could draw with people's information-consuming habits in other areas. But that's a problem for another day...)

Only You Can Prevent Your Mind From Getting Killed By Politics

Follow-up to: "Politics is the mind-killer" is the mind-killer, Trusting Expert Consensus

Gratuitous political digs are to be avoided. Indeed, I edited my post on voting to keep it from sounding any more partisan than necessary. But the fact that writers shouldn't gratuitously mind-kill their readers doesn't mean that, when they do, the readers' reaction is rational. The rules for readers are different from the rules for writers. And it *especially* doesn't mean that when a writer talks about a "political" topic for a reason, readers can use "politics!" as an excuse for attacking a statement of fact that makes them uncomfortable.

Imagine an alternate history where Blue and Green remain important political identities into the early stages of the space age. Blues, for complicated ideological reasons, tend to support trying to put human beings on the moon, while Greens, for complicated ideological reasons, tend to oppose it. But in addition to the ideological reasons, it has become popular for Greens to oppose attempting a moonshot on the grounds that the moon is made of cheese, and any landing vehicle put on the moon would sink into the cheese.

Suppose you're a Green, but you know perfectly well that the claim the moon is made of cheese is ridiculous. You tell yourself that you needn't be too embarrassed by your fellow Greens on this point. On the whole, the Green ideology is vastly superior to the Blue ideology, and furthermore some Blues have begun arguing we should go to the moon because the moon is made of gold and we could get rich mining the gold. That's just as ridiculous as the assertion that the moon is made of cheese.

Now imagine that one day, you're talking with someone who you *strongly suspect* is a Blue, and they remark on how irrational it is for so many people to believe the moon is made of cheese. When you hear that, you may be inclined to get defensive. Politics is the mind-killer, arguments are soldiers, so the point about the irrationality of the cheese-mooners may suddenly sound like a soldier for the other side that must be defeated.

Except... you *know* the claim that the moon is made of cheese is ridiculous. So let me suggest that, in that moment, it's your duty as a rationalist to *not* chastise them for making such a "politically charged" remark, and *not* demand they refrain from saying such things unless they make it *perfectly clear* they're not attacking all Greens or saying it's irrational to oppose a moon shot, or anything like that.

Quoth Eliezer:

> Robin Hanson recently proposed stores where banned products could be sold. There are a number of excellent arguments for such a policy—an inherent right of individual liberty, the career incentive of bureaucrats to prohibit everything, legislators being just as biased as individuals. But even so (I replied), some poor, honest, not overwhelmingly educated mother of 5 children is going to go into these stores and buy a "Dr. Snakeoil's Sulfuric Acid Drink" for her arthritis and die, leaving her orphans to weep on national television.

> I was just making a simple factual observation. Why did some people think it was an argument in favor of regulation?

Just as commenters shouldn't have assumed Eliezer's factual observation was an argument in favor of regulation, you shouldn't assume the suspected Blue's observation is a pro-moon shot or anti-Green argument.

The above parable was inspired by some of the discussion of global warming I've seen on LessWrong. According to the 2012 LessWrong readership survey, the mean confidence of LessWrong readers in human-caused global warming is 79%, and the *median* confidence is 90%. That's more or less in line with the current scientific consensus.

Yet references to anthropogenic global warming (AGW) in posts on LessWrong often elicit negative reactions. For example, last year Stuart Armstrong once wrote a post titled, "Global warming is a better test of irrationality than theism." His thesis was non-obvious, yet on reflection, I think, probably correct. AGW-denialism is a closer analog to creationism than theism. As bad as theism is, it isn't a rejection of a generally accepted (among scientists) scientific claim with a lot of evidence behind it just because the claim clashes with your ideological. Creationism and AGW-denialism do fall under that category, though.

Stuart's post was massively down voted—currently at -2, but at one point I think it went as low as -7. Why? Judging from the comments, not because people were saying, "yeah, global warming denialism is irrational, but it's not clear it's worse than theism." Here's the most-upvoted comment (currently at +44), which was also cited as "best reaction I've seen to discussion of global warming *anywhere*" in the comment thread on my post Trusting Expert Consensus:

> Here's the main thing that bothers me about this debate. There's a set of many different questions involving the degree of past and current warming, the degree to which such warming should be attributed to humans, the degree to which future emissions would cause more warming, the degree to which future emissions will happen given different assumptions, what good and bad effects future warming can be expected to have at different times and given what assumptions (specifically, what probability we should assign to catastrophic and even existential-risk damage), what policies will mitigate the problem how much and at what cost, how important the problem is relative to other problems, what ethical theory to use when deciding whether a policy is good or bad, and how much trust we should put in different aspects of the process that produced the standard answers to these questions and alternatives to the standard answers. These are questions that empirical evidence, theory, and scientific authority bear on to different degrees, and a LessWronger ought to separate them out as a matter of habit, and yet even here some vague combination of all these questions tends to get mashed together into a vague question of whether to believe "the global warming consensus" or "the pro-global warming side", to the point where when Stuart says some class of people is more irrational than theists, I have no idea if he's talking about me. If the original post had said something like, "everyone whose median estimate of climate sensitivity to doubled CO2 is lower than 2 degrees Celsius is more irrational than theists", I might still complain about it falling afoul of anti-politics norms, but at least it would help create the impression that the debate was about ideas rather than tribes.

If you read Stuart's original post, it's clear this comment is reading ambiguity into the post where none exists. You could argue that Stuart was a *little* careless in switching between talking about AGW and global warming simpliciter, but I think his meaning is clear: he thinks rejection of AGW is irrational, which entails that he thinks the stronger "no warming for any reason" claim is irrational. And there's *no justification whatsoever* for suggesting Stuart's post could be read as saying, "if your estimate of future warming is only 50% of the estimate I prefer you're irrational"—or as taking a position on ethical theories, for that matter.

What's going on here? Well, the LessWrong readership is mostly on-board with the scientific view on global warming. But many identify as libertarians, and they're aware that in the US many *other* conservatives/libertarians reject that scientific consensus (and no, that's not just a stereotype). So hearing someone say AGW denialism is irrational is really uncomfortable for them, *even if they agree.* This leaves them wanting *some kind of excuse to* complain, one guy thinks of "this is ambiguous and too political" as that excuse, and a bunch of people upvote it.

(If you still don't find any of this odd, think of the "skeptic" groups that freely mock ufologists or psychics or whatever, but which are reluctant to say anything bad about religion, even though in truth the group is dominated by atheists. Far from a perfect parallel, but it's still worth thinking about.)

When the title for this post popped into my head, I had to stop and ask myself if it was actually true, or just a funny Smokey the Bear reference. But in an important sense it is: the broader society isn't going to stop spontaneously labeling various straightforward empirical questions as Blue or Green issues. If you want to stop your mind from getting killed by whatever issues other people have decided are political, the only way is to control how you react to that.

Systematic Lucky Breaks

Many people can point to significant events that improved their lives in a positive way. They often refer to these as "lucky breaks", and take it for granted that such events are rare. But most of the time "lucky breaks" don't need to be uncommon-you can often reverse engineer the reasons behind them and cause them to happen more frequently. So when a one-off event ends up contributing a lot of value, you should systematically make it part of your life.

Example 1: in June the Less Wrong - Cambridge community held a mega-meetup with several people arriving from out of state. Since several of us had to stay up until 2AM+ in order to meet with people, we decided to have a game night that evening, which I held at my place. The game night was excellent-plenty of people showed up, we all had a lot of fun, and it was a great way to socialize with several people. Since it went so well, I started hosting game nights regularly, eventually converging on one game night every two weeks. This was a phenomenal move in many ways-it let me meet a lot of interesting people, deepen my connections with my friends, quickly integrate with the Less Wrong community, and just in general have a lot of fun, simply by taking one thing that worked well and making it systematic.

Example 2: a while back I was given an assignment to set up a scalable analytic architecture to allow data scientists to iterate faster-a project where I had no idea what to do or how to start. In desperation, I reached out to several people on LinkedIn who had experience with similar projects. Some of them responded, and the advice I got was incredibly valuable, easily shaving months off of my learning curve. But there is no reason for me to only do this when I am completely desperate. Thus I've continued to reach out to experts when I have new projects, and this has allowed me to avoid mistakes and solve new problems much more quickly. This has significantly improved my learning speed and made a qualitative difference in how I work. I no longer dismiss potential ideas simply because I have no idea how to implement them-instead, I now talk to experts and figure out roughly how difficult those ideas are, which has allowed me to solve several problems I would have dismissed as unfeasibly difficult before.

Example 3: a few years back some of my friends in the tech industry mentioned that Machine Learning was becoming a trend, so I took two weeks to learn the basics. A few months later the "Big Data" boom exploded, and I was able to get a job as a Data Scientist at a significantly higher salary doing more interesting work. Even though my Machine Learning knowledge was pretty rudimentary, I was able to get the job because demand completely exceeded supply at that point. In short, this was a lucky break that greatly advanced my career. To systematize this I simply continued to keep an eye out on big trends in technology. I've read Hacker News (which is generally half a year or more ahead of the mainstream), kept in touch with my friends on the applied side of academia (which feeds useful techniques into the industry), and just generally kept talking to a lot of people in order to keep up-to-date. This has been useful again

and again, allowing me to focus my learning on the most valuable skills right as there was market demand.

In short, one of the fastest ways to improve your life is to look at things that already made a big difference before, and cause more of them to happen.

Does Goal Setting Work?

tl;dr There's some disagreement over whether setting goals is a good idea. Anecdotally, enjoyment in setting goals and success at accomplishing them varies between people, for various possible reasons. Publicly setting goals may reduce motivation by providing a status gain before the goal is actually accomplished. Creative work may be better accomplished without setting goals about it. 'Process goals', 'systems' or 'habits' are probably better for motivation than 'outcome' goals. Specific goals are probably easier on motivation than unspecified goals. Having explicit set goals can cause problems in organizations, and maybe for individuals.

Introduction

> I experimented by letting go of goals for a while and just going with the flow, but that produced even worse results. I know some people are fans of that style, but it hasn't worked well for me. I make much better progress — and I'm generally happier and more fulfilled — when I wield greater conscious control over the direction of my life.

Steve Pavlina

> The inherent problem with goal setting is related to how the brain works. Recent neuroscience research shows the brain works in a protective way, resistant to change. Therefore, any goals that require substantial behavioural change or thinking-pattern change will automatically be resisted. The brain is wired to seek rewards and avoid pain or discomfort, including fear. When fear of failure creeps into the mind of the goal setter it commences a de-motivator with a desire to return to known, comfortable behaviour and thought patterns.

Ray Williams

I can't read these two quotes side by side and not be confused.

There's been quite a bit of discussion within Less Wrong and CFAR about goals and goal setting. On the whole, CFAR seems to go with it being a good idea. There are some posts that recognize the possible dangers: see patrissimo's post on the problems with receiving status by publicly committing to goals. Basically, if you can achieve the status boost of actually accomplishing a goal by just talking about it in public, why do the hard work? This discussion came up fairly recently with the Ottawa Less Wrong group; specifically, whether introducing group goal setting was a good idea.

I've always set goals–by 'always' I mean 'as far back as I can identify myself as some vaguely continuous version of my current self.' At age twelve, some of my goals were concrete and immediate–"get a time under 1 minute 12 seconds for a hundred freestyle and make the regional swim meet cut." Some were ambitious and unlikely–"go to the Olympics for swimming," and "be the youngest person to swim across Lake Ontario." Some were vague, like "be beautiful" or "be a famous novelist." Some were chosen for bad reasons, like "lose 10 pounds." My 12-year-old self wanted

plenty of things that were unrealistic, or unhealthy, or incoherent, but I wanted them, and it seemed to make perfect sense to do something about getting them. I took the bus to swim practice at six am. I skipped breakfast and threw out the lunch my mom packed. Et cetera. I didn't write these goals down in a list format, but I certainly kept track of them, in diary entries among other things. I sympathize with the first quote, and the second quote confuses and kind of irritates me–seriously, Ray Williams, you have that little faith in people's abilities to change?

For me personally, I'm not sure what the alternative to having goals would be. Do things at random? Do whatever you have an immediate urge to do? Actually, I do know people like this. I know people whose stated desires aren't a good predictor of their actions at all, and I've had a friend say to me "wow, you really do plan everything. I just realized I don't plan anything at all." Some of these people get a lot of interesting stuff done. So this may just be an individual variation thing; my comfort with goal setting, and discomfort with making life up as I go, might be a result of my slightly-Aspergers need for control. It certainly comes at a cost–the cost of basing self-worth on an external criterion, and the resulting anxiety and feelings of inadequacy. I have an enormous amount of difficulty with the Buddhist virtue of 'non-striving.'

Why the individual variation?

The concepts of the motivation equation and success spirals give another hint at why goal-driven behaviour might vary between people. Nick Winter talks about this in his book The Motivation Hacker; he shows the difference between his past self, who had very low expectancy of success and set few goals, and his present self, with high expectancy of success and with goal-directed behaviour filling most of his time.

I actually remember a shift like this in my own life, although it was back in seventh grade and I've probably editorialized the memories to make a good narrative. My sixth grade self didn't really have a concept of wanting something and thus doing something about it. At some point, over a period of a year or two, I experienced some minor successes. I was swimming faster, and for the first time ever, a coach made comments about my 'natural talent.' My friends wanted to get on the honour roll with an 80% average, and in first semester, both of them did and I didn't; I was upset and decided to work harder, a concept I'd never applied to school, and saw results the next semester when my average was on par with theirs. It only took a few events like that, inconsequential in themselves, before my self-image was of someone who could reliably accomplish things through hard work. My parents helpfully reinforced this self-stereotype by making proud comments about my willpower and determination.

In hindsight I'm not sure whether this was a defining year; whether it actually made the difference, in the long run, or whether it was inevitable that some cluster of minor successes would have set off the same cascade later. It may be that some innate personality trait distinguishes the people who take those types of experiences and interpret them as success spirals from those who remained disengaged.

The More Important Question

Apart from the question of personal individual variation, though, there's a more relevant question. Given that you're already at a particular place on the continuum from planning-everything to doing-everything-as-you-feel-like-it, how much should you

want to set goals, versus following urges? More importantly, what actions are helped versus harmed by explicit goal-setting.

Creative Goals

As Paul Graham points out, a lot of the cool things that have been accomplished in the past weren't done through self-discipline:

> One of the most dangerous illusions you get from school is the idea that doing great things requires a lot of discipline. Most subjects are taught in such a boring way that it's only by discipline that you can flog yourself through them. So I was surprised when, early in college, I read a quote by Wittgenstein saying that he had no self-discipline and had never been able to deny himself anything, not even a cup of coffee.
>
> Now I know a number of people who do great work, and it's the same with all of them. They have little discipline. They're all terrible procrastinators and find it almost impossible to make themselves do anything they're not interested in. One still hasn't sent out his half of the thank-you notes from his wedding, four years ago. Another has 26,000 emails in her inbox.
>
> I'm not saying you can get away with zero self-discipline. You probably need about the amount you need to go running. I'm often reluctant to go running, but once I do, I enjoy it. And if I don't run for several days, I feel ill. It's the same with people who do great things. They know they'll feel bad if they don't work, and they have enough discipline to get themselves to their desks to start working. But once they get started, interest takes over, and discipline is no longer necessary.
>
> Do you think Shakespeare was gritting his teeth and diligently trying to write Great Literature? Of course not. He was having fun. That's why he's so good.

This seems to imply that creative goals aren't a good place to apply goal setting. But I'm not sure how much this is a fundamental truth. I recently made a Beeminder goal for writing fiction, and I've written fifty pages since then. I actually don't have the writer's virtue of just sitting down and writing; in the past, I've written most of my fiction by staying up late in a flow state. I can't turn this on and off, though, and more importantly, I have a life to schedule my writing around, and if the only way I can get a novel done is to stay up all night before a 12-hour shift at the hospital, I probably won't write that novel. I rarely want to do the hard work of writing; it's a lot easier to lie in bed thinking about that one awesome scene five chapters down the road and lamenting that I don't have time to write tonight because work in the morning.

Even if Shakespeare didn't write using discipline, I bet that he used habits. That he sat down every day with a pen and parchment and fully expected himself to write. That he had some kind of sacred writing time, not to be interrupted by urgent-but-unimportant demands. That he'd built up some kind of success spiral around his ability to write plays that people would enjoy.

Outcome versus process goals

> Goal setting sets up an either-or polarity of success. The only true measure can either be 100% attainment or perfection, or 99% and less, which is failure. We can

> then excessively focus on the missing or incomplete part of our efforts, ignoring the successful parts. Fourthly, goal setting doesn't take into account random forces of chance. You can't control all the environmental variables to guarantee 100% success.
>
> Ray Williams

This quote talks about a type of goal that I don't actually set very often. Most of the ‘bad’ goals that I had as a 12-year-old were unrealistic outcome goals, and I failed to accomplish plenty of them; I didn’t go to the Olympics, I didn’t swim across Lake Ontario, and I never got down to 110 pounds. But I still have the self-concept of someone who’s good at accomplishing goals, and this is because I accomplished almost all of my more implicit ‘process’ goals. I made it to swim practice seven times a week, waking up at four-thirty am year after year. This didn’t automatically lead to Olympic success, obviously, but it was hard, and it impressed people. And yeah, I missed a few mornings, but in my mind 99% success or even 90% success at a goal is still pretty awesome.

In fact, I can’t think of any examples of outcome goals that I’ve set recently. Even “become a really awesome nurse” feels like more of a process goal, because it's something I'll keep doing on a day-to-day basis, requiring a constant input of effort.

Scott Adams, of Dilbert fame, refers to this dichotomy as ‘systems’ versus ‘goals’:

> Just after college, I took my first airplane trip, destination California, in search of a job. I was seated next to a businessman who was probably in his early 60s. I suppose I looked like an odd duck with my serious demeanor, bad haircut and cheap suit, clearly out of my element. I asked what he did for a living, and he told me he was the CEO of a company that made screws. He offered me some career advice. He said that every time he got a new job, he immediately started looking for a better one. For him, job seeking was not something one did when necessary. It was a continuing process... This was my first exposure to the idea that one should have a system instead of a goal. The system was to continually look for better options.
>
> Throughout my career I've had my antennae up, looking for examples of people who use systems as opposed to goals. In most cases, as far as I can tell, the people who use systems do better. The systems-driven people have found a way to look at the familiar in new and more useful ways.
>
> …To put it bluntly, goals are for losers. That's literally true most of the time. For example, if your goal is to lose 10 pounds, you will spend every moment until you reach the goal—if you reach it at all—feeling as if you were short of your goal. In other words, goal-oriented people exist in a state of nearly continuous failure that they hope will be temporary.
>
> If you achieve your goal, you celebrate and feel terrific, but only until you realize that you just lost the thing that gave you purpose and direction. Your options are to feel empty and useless, perhaps enjoying the spoils of your success until they bore you, or to set new goals and re-enter the cycle of permanent presuccess failure.

I guess I agree with him–if you feel miserable when you've lost 9 pounds because you haven't accomplished your goal yet, and empty after you've lost 10 pounds because you no longer have a goal, then whatever you're calling 'goal setting' is a terrible idea.

But that's not what 'goal setting' feels like to me. I feel increasingly awesome as I get closer towards a goal, and once it's done, I keep feeling awesome when I think about how I did it. Not awesome enough to never set another goal again, but awesome enough that I want to set lots more goals to get that feeling again.

SMART goals

> When I work with people as their coach and mentor, they often tell me they've set goals such as "I want to be wealthy," or "I want to be more beautiful/popular," "I want a better relationship/ideal partner." They don't realize they've just described the symptoms or outcomes of the problems in their life. The cause of the problem, that many resist facing, is themselves. They don't realize that for a change to occur, if one is desirable, they must change themselves. Once they make the personal changes, everything around them can alter, which may make the goal irrelevant.

Ray Williams

And? Someone has to change themselves to fix the underlying problem? Are they going to do that more successfully by going with the flow?

I think the more important dichotomy here is between vague goals and specific goals. I was exposed to the concept of SMART goals (specific, measurable, attainable, relevant, time-bound), at an early age, and though the concept has a lot of problems, the ability to Be Specific seems quite important. You can break down “I want to be beautiful” into subgoals like “I’ll learn to apply makeup properly”, “I’ll eat healthy and exercise”, “I’ll go clothing shopping with a friend who knows about fashion,” etc. All of these feel more attainable than the original goal, and it’s clear when they’re accomplished.

That being said, I have a hard time setting any goal that isn’t specific, attainable, and small. I’ve become more ambitious since meeting lots of LW and CFAR people, but I still don’t like large, long-term goals unless I can easily break them down into intermediate parts. This makes the idea of working on an unsolved problem, or in a startup where the events of the next year aren’t clear, deeply frightening. And these are obviously important problems that someone needs to motivate themselves to work on.

Problematic Goal-Driven Behaviour

> We argue that the beneficial effects of goal setting have been overstated and that systematic harm caused by goal setting has been largely ignored. We identify specific side effects associated with goal setting, including a narrow focus that neglects non-goal areas, a rise in unethical behaviour, distorted risk preferences, corrosion of organizational culture, and reduced intrinsic motivation. Rather than dispensing goal setting as a benign, over-the-counter treatment for motivation, managers and scholars need to conceptualize goal setting as a prescription-strength medication that requires careful dosing, consideration of harmful side effects, and close supervision.

Goals Gone Wild

This is a fairly compelling argument against goal-setting; that by setting an explicit goal and then optimizing towards that goal, you may be losing out on elements that were being accomplished better before, and maybe even rewarding actual negative behaviour. Members of an organization presumably already have assigned tasks and responsibilities, and aren't just doing whatever they feel like doing, but they might have done better with more freedom to prioritize their own work–the best environment is one with some structure and goals, but not too many. The phenomenon of "teaching to the test" for standardized testing is another example.

Given that humans aren't best described as unitary selves, this metaphor extends to individuals. If one aspect of myself sets a personal goal to write two pages per day, another aspect of myself might respond by writing two pages on the easiest project I can think of, like a journal entry that no one will ever see. This violates the spirit of the goal it technically accomplishes.

A more problematic consideration is the relationship between intrinsic and extrinsic motivation. Studies show that rewarding or punishing children for tasks results in less intrinsic motivation, as measured by stated interest or by freely choosing to engage in the task. I've noticed this tendency in myself; faced with a nursing instructor who was constantly quizzing me on the pathophysiology of my patients' conditions, I responded by refusing to be curious about any of it or look up the answers to questions in any more detail than what she demanded, even though my previous self loved to spend hours on Google making sense of confusing diseases. If this is a problem that affects individuals setting goals for themselves–i.e. if setting a daily writing goal makes writing less fun–then I can easily see how goal-setting could be damaging.

I also notice that I'm confused about the relationship between Beeminder's extrinsic motivation, in the form of punishment for derailing, and its effects on intrinsic motivation. Maybe the power of success spirals to increase intrinsic motivation offsets the negative effect of outside reward/punishment; or maybe the fact that users deliberately choose to use Beeminder means that it doesn't count as "extrinsic." I'm not sure.

Conclusion

There seems to be variation between individuals, in terms of both generally purposeful behaviour, and comfort level with calling it 'setting goals'. This might be related to success spirals in the past, or it might be a factor of personality and general comfort with order versus chaos. I'm not sure if it's been studied.

In the past, a lot of creative behaviour wasn't the result of deliberate goals. This may be a fundamental fact about creativity, or it may be a result of people's beliefs about creativity (à la ego depletion only happens if you belief in ego depletion) or it may be a historical coincidence that isn't fundamental at all. In any case, if you aren't currently getting creative work done, and want to do more, I'm not sure what the alternative is to purposefully trying to do more. Manipulating the environment to make flow easier to attain, maybe. (For example, if I quit my day job and moved to a writers' commune, I might write more without needing to try on a day-to-day basis).

Process goals, or systems, are probably better than outcome goals. Specific and realistic goals are probably better than vague and ambitious ones. A lot of this may be

because it’s easier to form habits and/or success spirals around well-specified behaviours that you can just do every day.

Setting goals within an organization has a lot of potential problems, because workers can game the system and accomplish the letter of the goal in the easiest possible way. This likely happens within individuals too. Research shows that extrinsic motivation reduces intrinsic motivation, which is important to consider, but I'm not sure how it relates to individuals setting goals, as opposed to organizations.

Meditation Trains Metacognition

Summary: Some forms of meditation may train key skills of metacognition, serving as powerful tools for applied rationality. I expect aspiring rationalists to advance more quickly with a regular practice of mindfulness meditation.

The state of scientific research on meditation isn't great. Although there's evidence that it does something good--probably something involving down-regulation of negative affect--there are many basic questions[1] that either haven't been studied at all or haven't been studied well enough to let me update much. According to a meta-analysis by Sedlmeier et al., one problem with evaluating the research is that it's hard to pin down what meditation is, let alone what it does or why it does it. In their words,

> ...two of our main findings are that (a) meditation has a substantial impact on psychological variables, indicated by a medium-sized (e.g., Cohen, 1988) global effect, and (b) its effects might be somewhat stronger for negative emotional than for cognitive variables. Due to the lack of a comprehensive theoretical approach (and results from studies derived therefrom), it is still unclear how meditation works... Moreover, a closer look at the studies included in the meta-analysis revealed that they differed in many respects that might have affected the results.[2]

So I just want to be clear that I don't mean in this post to wholeheartedly recommend daily meditation as the best possible use of 1/24th of your time.

Nevertheless, my own experience and reports from several of my friends suggest a specific cognitive result from a certain flavor of meditation that will be very good news for rationality if we can reliably reproduce it.[3] In a recent post, Julia Galef pointed out exactly what I consider to be far and away the greatest benefit I've reaped from my meditative practices over the years. She wrote,

> Meditation seems to train you to stop automatically identifying with all of your thoughts, so that, for example, when the thought "John's a jerk" pops into your head, you don't assume that John necessarily is a jerk. You take the thought as something your brain produced, which may or may not be true, and may or may not be useful -- and this ability to take a step back from your thoughts and reflect on them is arguably one of the building blocks of rationality.

I'd like to delve more deeply into how and why this could work. There seem to be multiple paths to establishing the central rationality skills comprising metacognition--several highly advanced rationalist I know have no background in meditation--so meditation is by no means a necessary condition for successfully applied rationality. I think it may, however, have the highest signal to noise ratio among methods for developing foundational metacognitive abilities. At a minimum, I expect that regular practice of certain kinds of meditation would help aspiring rationalists to advance more quickly.

What kind of metacognitive skills am I talking about?

How about an example. When you read these words, you're probably hearing a little voice inside your head that's reading them to you aloud, so to speak. Your relationship to this imaginary voice (aka "subvocalization" or "inner speech") may be quite a bit more intimate than you realize. It's likely with you not only when you read, but when you ride the bus home and think, "Maybe I'll have steak for dinner"; it's with you when you've had an awkward interaction with someone you admire and you think, "God, I must have looked like such an idiot"; it's with you most of the time, in fact, during your waking hours and maybe even when you dream.[5]

Exercise One

This fact may be more salient for you if you try to turn it off for a while. Set a timer for one minute, and force yourself not to verbally narrate your experience. When the minute is up, jot down a brief note about how it felt. Three two one go.

No, really, don't read the next paragraph 'til you've done the exercise.[6]

Even if you managed to go the entire minute without subvocalizing, it probably didn't feel like the natural way of things. It probably took effort, and possibly a great deal of effort. But I predict that most of you didn't go through the whole minute in total subjective silence. (If you succeeded and it did feel like the natural way of things, I'd very much like to know.)

Exercise Two

Now set the timer for a minute again, but this time *don't* force yourself not to subvocalize. Simply notice when words arise in consciousness. Don't bother doing anything with them. Just be aware of them.

Again, when the minute is up, make a brief note about how it felt. In particular, how did it differ from the first exercise, and how did it differ from your usual experience?

Exercise Three

Finally, notice that your present increased awareness of subvocalizations lets you change things about them that you couldn't change if you weren't aware of them. For example, you're now reading this in the voice of Morgan Freeman. (You're welcome.)

Now pick some other aspect of subvocalization to change--perhaps the accent, or the speed, or the pitch--and read the next sentence in that way. Set a timer for one minute, and experiment with things you can change about your experience of inner speech.

What does this have to do with rationality?

In general terms, what have you done in the above exercises? You've become aware of a mental process that usually runs in the background whether you like it or not. You've gained and exercised some degree of control over it. You've come up with and tested, in real time, alternative ways of running your own cognitive software.

Now, this has merely been a simple illustration. My point is not that swapping Morgan Freeman for yourself as official narrator would itself improve your daily life. (Although that may well be.) Rather, my point is that these skills are central to rationality and are cultivated by meditation. Those of you with a strong background in meditation

probably did not learn anything important from these exercises, and wouldn't have regardless of your rationality training. Stepping back from your experiences in a way that lets you examine them and modify them is so old hat, if you meditate a lot, that you may even have forgotten what it's like not to have that action available as primitive.

This is extraordinarily valuable! There are three abilities that together form the bridge between knowledge of rationality and the application thereof. They are

1. the ability to introspect and promote a sought cognitive process into consciousness
2. the ability to not identify with any particular cognitive process you become aware of
3. the ability to make changes to cognitive processes you're aware of in media res

For example, even if you understand how important it is to make beliefs pay rent in anticipated experiences, actually doing that can be really hard. Why is it so hard? Possibly for a few different reasons, but prominent among them is the following. If you've thought something lots of times without ever explicitly identifying it as something you're thinking, without putting much distance between yourself and the thought, your sense of self gets tangled up in it. It's not nice to let go of something that close to you, even if it's useless or harmful. It feels sort of like trying to kick out your own child when you know you can no longer afford to take care of her--and it feels distinctly unlike taking a broken blender to the dump, which is closer to what should really be going on.

Other directly related examples include noticing and tending to confusion, actually behaving as though you might be wrong when you think you might be wrong, and thinking about politics without your head exploding.

Note that merely willing problems solved is not a reliable way of solving them. Resolving to not identify with your thoughts isn't the same as *causing* yourself to not identify with your thoughts. There's a reason you identify with your thoughts in the first place, and it's not because you decided to. If you don't alter any of the mechanisms that actually give rise to the problem, nothing will change--which is why, I think, it's possible to possess oodles of declarative knowledge about rationality without making a single significant improvement to your life.

One way or another, you have to get some distance between yourself and your thoughts and feelings if you want to let go of them or change them. That's exactly what meditation teaches you to do.

What does this have to do with meditation?

There are many kinds of meditation. Some involve intense concentration on very specific sensations, like visualizations of geometric patterns, repeated phrases called mantra, the breath, or movements (not all meditation is done seated and motionless). There are interpersonal types of meditation that can involve maintaining eye contact with someone for extended periods, imagining someone hurting and nurturing the desire to help them, or sex. The kind I'm most familiar with is a form of Japanese Buddhist meditation called shikantaza, which translates roughly to "just sitting". Although it comes with basically no instructions, as the name suggests, in practice it's nearly identical to the most general form of "mindfulness meditation".

Mindfulness is one of the most popular meditative practices in the West, and of the types I know about, it's the one I expect to be most relevant to applied rationality. Though all of the above, in one way or another, teach the backward step[7] that allows you to stop identifying with thoughts, mindfulness is *only* that. Exercise two above is a limited form of mindfulness meditation. Although there's a whole family of practices that fall under the heading of "mindfulness", what they have in common is the cultivation of awareness.

All I mean by "cultivation of awareness" is the power to broaden/focus attention to encompass more things, or more specific things. I've often heard practitioners describe it as "openness to the world". Ordinarily, we experience a lot of things on which we don't bother to turn our subjective spotlights of attention, sometimes because they're just not important, and sometimes because we actively avoid stimuli we perceive to be aversive.

Subvocalization is an example. Other examples are the sounds in your external environment, what you know about how those around you are feeling, the sound of your own breath and heartbeat, the sensation of flinching away from a painful thought, the temperature in the room, the colors and shapes that appear behind closed eyelids, and the sensation of confusion. I find it difficult to describe the most general form of this, because without analogy to more specific forms, all I've got is that it's experiencing... what you experience. Which really just sounds like the default mode of living, doesn't it? But in practice it can feel very different.

When you're well practiced at noticing these things, at welcoming them into your attention, you're acutely aware of not being them. And when you don't feel as though you are your thoughts and feelings, it becomes emotionally easier to let go of them or to modify them. Changing your mind feels less like losing a part of yourself.

Further Resources

- Sam Harris recently posted an excellent introduction to mindfulness meditation in the form of two audio tracks (one nine minutes, the other twenty-six). I recommend them pretty highly. They each guide you through a meditation session without any annoying religious or new-agey distractions.
- I find this meta-analysis by Northoff et al. of neuroimaging studies of self-referential cognitive processing to be fascinating for all sorts of reasons. Chief among them is the light it sheds on how and why including clear-cut self/other distinctions in models of human minds doesn't always work so well. (Link is to a PDF.)
- If you want to take a soaring leap across the bridge between knowledge of rationality and the application thereof, you simply must try a CFAR workshop. I did one of these back in April, and it was every bit as fun as it was effective (which was very).

Notes

1. Off the top of my head: What aspects of particular forms of meditation cause the various purported benefits? If we pinpoint those aspects, can we harness their corresponding benefits individually without committing to meditation as a whole? Can

we improve upon them? Does meditation have different effects when practiced in a religious context? What is the relationship between meditation and hypnosis? How do the effects differ among different age groups? Does learning to meditate while young have any effect on adult meditation?

2. Sedlmeier et al. (2012). The Psychological Effects of Meditation: A Meta-Analysis. Psychological Bulletin, 138(6) 1139-1171.

3. I don't consider lack of supporting double-blind studies much evidence against my thesis, largely because the result in question would only show up in tests of metacognitive techniques I expect not to occur to the vast majority of researchers just yet.

4. In case you're wondering about my relevant background: I did Vinyasa yoga throughout high school, taught it during college, trained at a residential Soto Zen temple for a summer, have maintained a fairly regular practice of Zen meditation for about five years now, practiced Tai Chi (very casually) off and on for most of my life, and have a degree in religious studies with a focus on East Asian Buddhism.

5. Fun fact: You internally simulate your voice in parallel with actual talking.

6. Yes, you're doing it right. If you're trying to do it at all, you're doing it right. The idea is to find out what it feels like to make the effort, not to beat the game. There is no game. Some of the comments below have me concerned that I may be contributing to the "meditation means being brain dead" misconception. This exercise isn't meant to teach you the One True Way of Meditation. It's just to point at certain kinds of movements your mind makes. Monks who have been meditating for multiple hours a day for decades don't have completely featureless minds when they meditate. That isn't even close to what meditation means to them. Beginners are given exercises along these lines because it's an easier entry point, like training wheels. Eventually, counting breaths simply becomes irrelevant.

7. The Japanese “Su sube[karaku] mochi[iyo] eko-hensho no taiho o mochi-iyo,” from Dogen Zengi’s instructions for meditation (1227) literally translates to English (character-by-character) as “Remember/employ of backward step turning light/consciousness reflecting/illuminating.” (Dogen loved wordplay, and the double meanings are intentional.) Though most translators render this something like, “Take the backward step that turns your light inwardly to illuminate the self,” the characters for “self” and “inward” do not in fact appear in this part of the text. Thus, his central instruction for meditation is to step consciousness backward so it can be generally reflective. This is why I say that in practice shikantaza and mindfulness amount to the same thing.