Best of LessWrong: June 2014

1. New organization - Future of Life Institute (FLI)
2. Willpower Depletion vs Willpower Distraction
3. [Meta] The Decline of Discussion: Now With Charts!
4. On Terminal Goals and Virtue Ethics
5. False Friends and Tone Policing
6. What resources have increasing marginal utility?
7. Against utility functions
8. Mathematics as a lossy compression algorithm gone wild
9. Flowers for Algernon

Best of LessWrong: June 2014

1. New organization - Future of Life Institute (FLI)
2. Willpower Depletion vs Willpower Distraction
3. [Meta] The Decline of Discussion: Now With Charts!
4. On Terminal Goals and Virtue Ethics
5. False Friends and Tone Policing
6. What resources have increasing marginal utility?
7. Against utility functions
8. Mathematics as a lossy compression algorithm gone wild
9. Flowers for Algernon

New organization - Future of Life Institute (FLI)

As of May 2014, there is an existential risk research and outreach organization based in the Boston area. The Future of Life Institute (FLI), spearheaded by Max Tegmark, was co-founded by Jaan Tallinn, Meia Chita-Tegmark, Anthony Aguirre and myself.

Our idea was to create a hub on the US East Coast to bring together people who care about x-risk and the future of life. FLI is currently run entirely by volunteers, and is based on brainstorming meetings where the members come together and discuss active and potential projects. The attendees are a mix of local scientists, researchers and rationalists, which results in a diversity of skills and ideas. We also hold more narrowly focused meetings where smaller groups work on specific projects. We have projects in the pipeline ranging from improving Wikipedia resources related to x-risk, to bringing together AI researchers in order to develop safety guidelines and make the topic of AI safety more mainstream.

Max has assembled an impressive advisory board that includes Stuart Russell, George Church and Stephen Hawking. The advisory board is not just for prestige - the local members attend our meetings, and some others participate in our projects remotely. We consider ourselves a sister organization to FHI, CSER and MIRI, and touch base with them often.

We recently held our launch event, a panel discussion "The Future of Technology: Benefits and Risks" at MIT. The panelists were synthetic biologist George Church, geneticist Ting Wu, economist Andrew McAfee, physicist and Nobel laureate Frank Wilczek and Skype co-founder Jaan Tallinn. The discussion covered a broad range of topics from the future of bioengineering and personal genetics, to autonomous weapons, AI ethics and the Singularity. A video and transcript are available.

FLI is a grassroots organization that thrives on contributions from awesome people like the LW community - here are some ways you can help:

- If you have ideas for research or outreach we could be doing, or improvements to what we're already doing, please let us know (in the comments to this post, or by contacting me directly).
- If you are in the vicinity of the Boston area and are interested in getting involved, you are especially encouraged to get in touch with us!
- Support in the form of donations is much appreciated. (We are grateful for seed funding provided by Jaan Tallinn and Matt Wage.)

More details on the ideas behind FLI can be found in this article.

Willpower Depletion vs Willpower Distraction

I once asked a room full of about 100 neuroscientists whether willpower depletion was a thing, and there was widespread disagreement with the idea. (A propos, this is a great way to quickly gauge consensus in a field.) Basically, for a while some researchers believed that willpower depletion "is" glucose depletion in the prefrontal cortex, but some more recent experiments have failed to replicate this, e.g. by finding that the mere taste of sugar is enough to "replenish" willpower faster than the time it takes blood to move from the mouth to the brain:

> Carbohydrate mouth-rinses activate dopaminergic pathways in the striatum–a region of the brain associated with responses to reward (Kringelbach, 2004)–whereas artificially-sweetened non-carbohydrate mouth-rinses do not (Chambers et al., 2009). Thus, the sensing of carbohydrates in the mouth appears to signal the possibility of reward (i.e., the future availability of additional energy), which could motivate rather than fuel physical effort.
>
> -- Molden, D. C. et al, The Motivational versus Metabolic Effects of Carbohydrates on Self-Control. Psychological Science.

Stanford's Carol Dweck and Greg Walden even found that hinting to people that using willpower is energizing might actually make them less depletable:

> When we had people read statements that reminded them of the power of willpower like, “Sometimes, working on a strenuous mental task can make you feel energized for further challenging activities,” they kept on working and performing well with no sign of depletion. They made half as many mistakes on a difficult cognitive task as people who read statements about limited willpower. In another study, they scored 15 percent better on I.Q. problems.
>
> -- Dweck and Walden, Willpower: It’s in Your Head? New York Times.

While these are all interesting empirical findings, there’s a very similar phenomenon that’s much less debated and which could explain many of these observations, but I think gets too little popular attention in these discussions:

Willpower is *distractible*.

Indeed, willpower and working memory are both strongly mediated by the dorsolateral prefontal cortex, so “distraction” could just be the two functions funging against one another. To use the terms of Stanovich popularized by Kahneman in *Thinking: Fast and Slow*, "System 2" can only override so many "System 1" defaults at any given moment.

So what’s going on when people say "willpower depletion"? I’m not sure, but even if willpower depletion is not a thing, the following distracting phenomena clearly are:

- Thirst
- Hunger
- Sleepiness
- Physical fatigue (like from running)

- Physical discomfort (like from sitting)
- That specific-other-thing you want to do
- Anxiety about willpower depletion
- Indignation at being asked for too much by bosses, partners, or experimenters...

... and "willpower depletion" might be nothing more than mental distraction by one of these processes. Perhaps it really is better to think of willpower as power (a rate) than energy (a resource).

If that's true, then figuring out what processes might be distracting us might be much more useful than saying "I'm out of willpower" and giving up. Maybe try having a sip of water or a bit of food if your diet permits it. Maybe try reading lying down to see if you get nap-ish. Maybe set a timer to remind you to call that friend you keep thinking about.

The last two bullets,

- Anxiety about willpower depletion
- Indignation at being asked for too much by bosses, partners, or experimenters...

are also enough to explain why being told willpower depletion isn't a thing might reduce the effects typically attributed to it: we might simply be less distracted by anxiety or indignation about doing "too much" willpower-intensive work in a short period of time.

Of course, any speculation about how human minds work in general is prone to the "typical mind fallacy". Maybe my willpower is depletable and yours isn't. But then that wouldn't explain why you can cause people to exhibit less willpower depletion by suggesting otherwise. But then again, most published research findings are false. But then again the research on the DLPFC and working memory seems relatively old and well established, and distraction is clearly a thing...

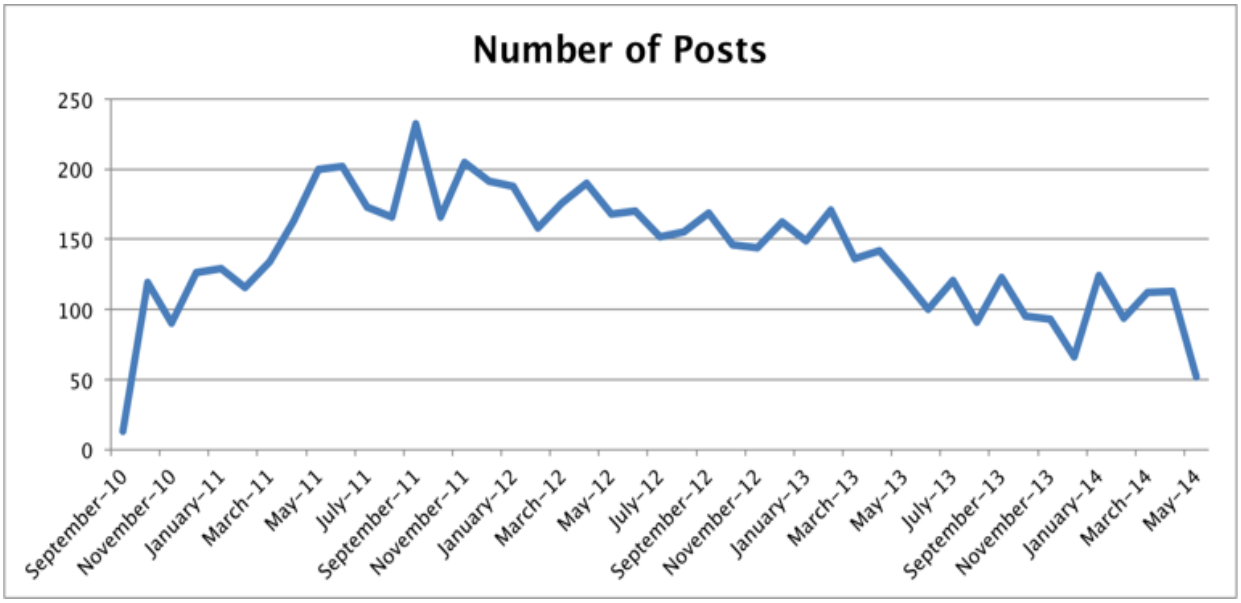
All in all, more of my chips are falling on the hypothesis that willpower "depletion" is often just willpower *distraction*, and that finding and addressing those distractions is probably a better a strategy than avoiding activities altogether in order to "conserve willpower".

[Meta] The Decline of Discussion: Now With Charts!

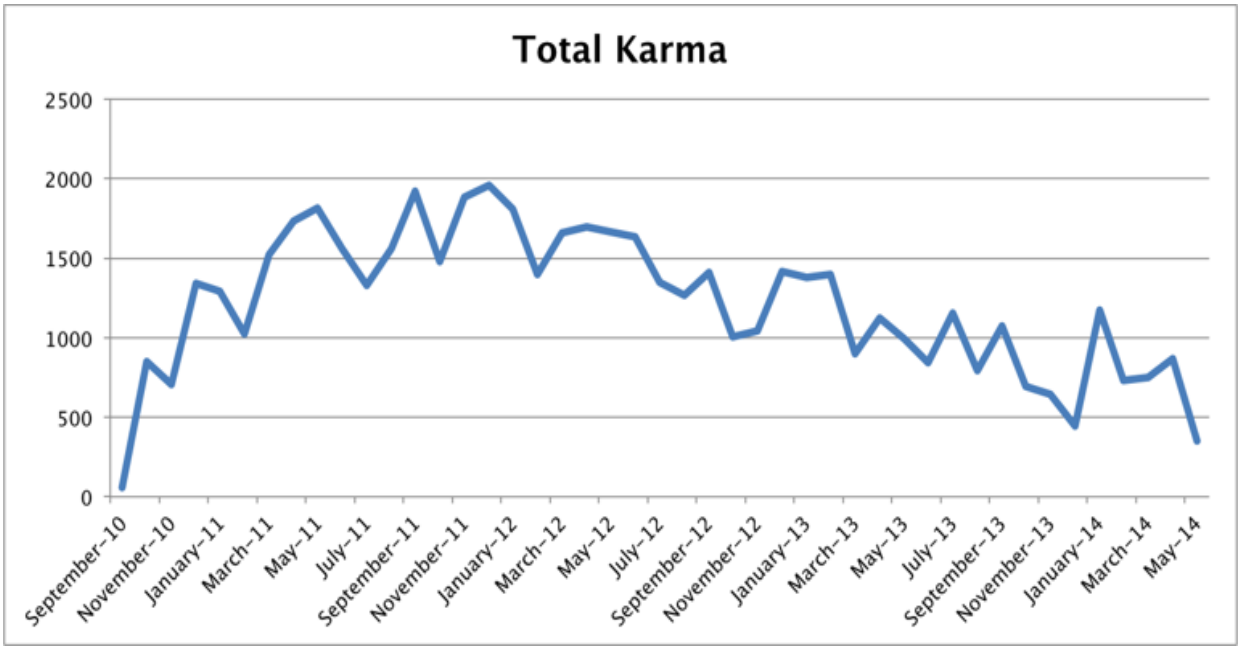
[Based on Alexandros's excellent dataset.]

I haven't done any statistical analysis, but looking at the charts I'm not sure it's necessary. The discussion section of LessWrong has been steadily declining in participation. My fairly messy spreadsheet is available if you want to check the data or do additional analysis.

Enough talk, you're here for the pretty pictures.

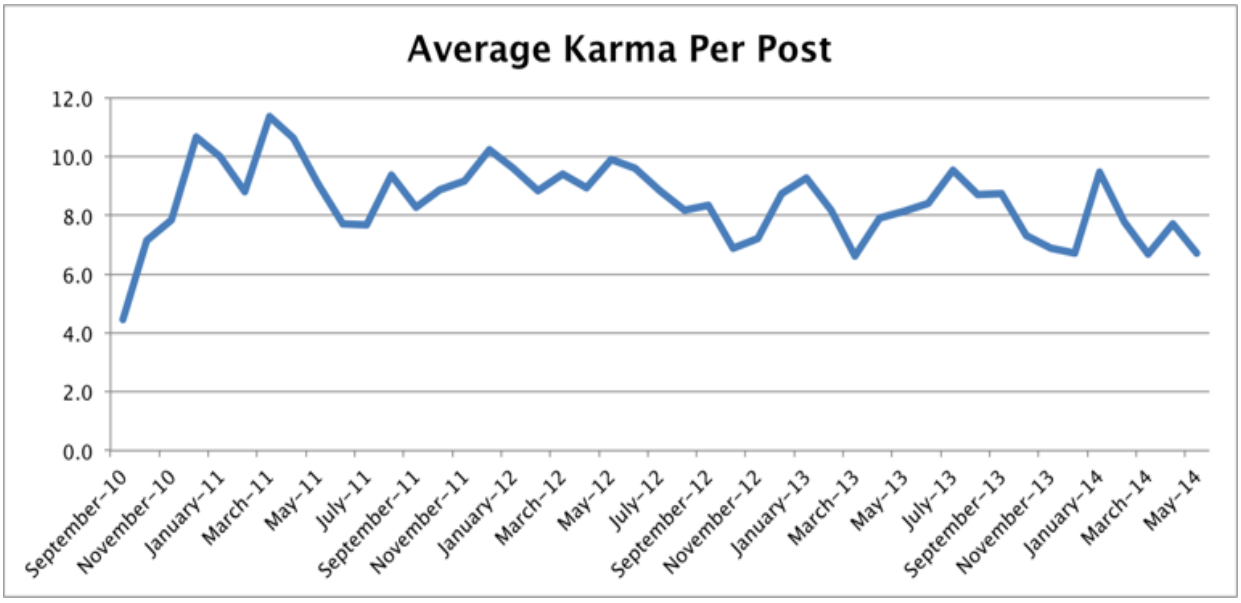


The number of posts has been steadily declining since 2011, though the trend over the last year is less clear. Note that I have excluded all posts with 0 or negative Karma from the dataset.



The total Karma given out each month has similarly been in decline.

Is it possible that there have been fewer posts, but of a higher quality?



No, at least under initial analysis the average Karma seems fairly steady. My prior here is that we're just seeing less visitors overall, which leads to fewer votes being distributed among fewer posts for the same average value. I would have expected the average karma to drop more than it did--to me that means that participation has dropped more steeply than mere visitation. Looking at the point values of the top posts would be helpful here, but I haven't done that analysis yet.

These are very disturbing to me, as someone who has found LessWrong both useful and enjoyable over the past few years. It raises several questions:

1. What should the purpose of this site be? Is it supposed to be building a movement or filtering down the best knowledge?
2. How can we encourage more participation?
3. What are the costs of various means of encouraging participation--more arguing, more mindkilling, more repetition, more off-topic threads, etc?

Here are a few strategies that come to mind:

Idea A: **Accept that LessWrong has fulfilled its purpose and should be left to fade away, or allowed to serve as a meetup coordinator and repository of the highest quality articles.** My suspicion is that without strong new content and an online community, the strength of the individual meetup communities may wane as fewer new people join them. This is less of an issue for established communities like Berkeley and New York, but more marginal ones may disappear.

Idea B: **Allow and encourage submission of rationalism, artificial intelligence, transhumanism etc related articles from elsewhere, possibly as a separate category.** This is how a site like Hacker News stays high engagement, even though many of the discussions are endless loops of the same discussion. It can be annoying for the old-timers, but new generations may need to discover things for themselves. Sometimes "put it all in one big FAQ" isn't the most efficient method of teaching.

Idea C: **Allow and encourage posts on "political" topics in Discussion (but probably NOT Main).** The dangers here might be mitigated by a ban on discussion of current politicians, governments, and issues. "Historians need to have had a decade to mull it over before you're allowed to introduce it as evidence" could be a good heuristic. Another option would be a ban on specific topics that cause the worst mindkilling. Obviously this is overall a dangerous road.

Idea D: **Get rid of Open Threads and create a new norm that a discussion post as short as a couple sentences is acceptable**. Open threads get stagnant within a day or two, and are harder to navigate than the discussion page. Moving discussion from the Open Threads to the Discussion section would increase participation if users could be convinced thatit was okay to post questions and partly-formed ideas there.

The challenge with any of these ideas is that they will require strong moderation.

At any rate, this data is enough to convince me that some sort of change is going to be needed in order to put the community on a growth trajectory. That is not necessarily the goal, but at its core LessWrong seems like it has the potential to be a powerful tool for the spreading of rational thought. We just need to figure out how to get it started into its next evolution.

On Terminal Goals and Virtue Ethics

Introduction

A few months ago, my friend said the following thing to me: “After seeing Divergent, I finally understand virtue ethics. The main character is a cross between Aristotle and you.”

That was an impossible-to-resist pitch, and I saw the movie. The thing that resonated most with me–also the thing that my friend thought I had in common with the main character–was the idea that you could make a particular decision, and set yourself down a particular course of action, in order to *make yourself become a particular kind of person.* Tris didn’t join the Dauntless cast because she thought they were doing the most good in society, or because she thought her comparative advantage to do good lay there–she chose it because they were brave, and she wasn’t, yet, and she wanted to be. Bravery was a virtue that she thought she ought to have. If the graph of her motivations even went any deeper, the only node beyond ‘become brave’ was ‘become good.’

(Tris did have a concept of some future world-outcomes being better than others, and wanting to have an effect on the world. But that wasn't the causal reason why she chose Dauntless; as far as I can tell, it was unrelated.)

My twelve-year-old self had a similar attitude. I read a lot of fiction, and stories had heroes, and I wanted to be like them–and that meant acquiring the right skills and the right traits. I knew I was terrible at reacting under pressure–that in the case of an earthquake or other natural disaster, I would freeze up and not be useful at all. Being good at reacting under pressure was an important trait for a hero to have. I could be sad that I didn’t have it, or I could decide to acquire it by doing the things that scared me over and over and over again. So that someday, when the world tried to throw bad things at my friends and family, I’d be *ready*.

You could call that an awfully passive way to look at things. It reveals a deep-seated belief that I’m not in control, that the world is big and complicated and beyond my ability to understand and predict, much less steer–that I am not the locus of control. But this way of thinking is an algorithm. It will almost always spit out an answer, when otherwise I might get stuck in the complexity and unpredictability of trying to make a particular outcome happen.

Virtue Ethics

I find the different houses of the HPMOR universe to be a very compelling metaphor. It’s not because they suggest actions to take; instead, they suggest virtues to focus on, so that when a particular situation comes up, you can act ‘in character.’ Courage and bravery for Gryffindor, for example. It also suggests the idea that different people can focus on different virtues–diversity is a useful thing to have in the world. (I'm

probably mangling the concept of virtue ethics here, not having any background in philosophy, but it's the closest term for the thing I mean.)

I've thought a lot about the virtue of loyalty. In the past, loyalty has kept me with jobs and friends that, from an objective perspective, might not seem like the optimal things to spend my time on. But the costs of quitting and finding a new job, or cutting off friendships, wouldn't just have been about direct consequences in the world, like needing to spend a bunch of time handing out resumes or having an unpleasant conversation. There would also be a shift within myself, a weakening in the drive towards loyalty. It wasn't that I thought *everyone* ought to be extremely loyal–it's a virtue with obvious downsides and failure modes. But it was a virtue that *I* wanted, partly because it seemed undervalued.

By calling myself a 'loyal person', I can aim myself in a particular direction without having to understand all the subcomponents of the world. More importantly, I can make decisions even when I'm rushed, or tired, or under cognitive strain that makes it hard to calculate through all of the consequences of a particular action.

Terminal Goals

The Less Wrong/CFAR/rationalist community puts a lot of emphasis on a different way of trying to be a hero–where you start from a terminal goal, like "saving the world", and break it into subgoals, and do whatever it takes to accomplish it. In the past I've thought of myself as being mostly consequentialist, in terms of morality, and this is a very consequentialist way to think about being a good person. And it doesn't feel like it would work.

There are some bad reasons why it might feel wrong–i.e. that it feels arrogant to think you can accomplish something that big–but I think the main reason is that it feels *fake.* There is strong social pressure in the CFAR/Less Wrong community to claim that you have terminal goals, that you're working towards something big. My System 2 understands terminal goals and consequentialism, as a thing that other people do–I could talk about my terminal goals, and get the points, and fit in, but I'd be lying about my thoughts. My model of my mind would be incorrect, and that would have consequences on, for example, whether my plans actually worked.

Practicing the art of rationality

Recently, Anna Salamon brought up a question with the other CFAR staff: "What is the thing that's wrong with your own practice of the art of rationality?" The terminal goals thing was what I thought of immediately–namely, the conversations I've had over the past two years, where other rationalists have asked me "so what are your terminal goals/values?" and I've stammered something and then gone to hide in a corner and try to come up with some.

In Alicorn's Luminosity, Bella says about her thoughts that "they were liable to morph into versions of themselves that were more idealized, more consistent - and not what

they were originally, and therefore false. Or they'd be forgotten altogether, which was even worse (those thoughts were mine, and I wanted them).”

I want to know true things about myself. I also want to impress my friends by having the traits that they think are cool, but not at the price of faking it–my brain *screams* that pretending to be something other than what you are isn’t virtuous. When my immediate response to someone asking me about my terminal goals is “but brains don’t work that way!” it may not be a true statement about all brains, but it’s a true statement about *my* brain. My motivational system is wired in a certain way. I could think it was broken; I could let my friends convince me that I needed to change, and try to shoehorn my brain into a different shape; or I could accept that it *works*, that I get things done and people find me useful to have around and *this is how I am.* For now. I'm not going to rule out future attempts to hack my brain, because Growth Mindset, and maybe some other reasons will convince me that it's important enough, but if I do it, it'll be on my terms. Other people are welcome to have their terminal goals and existential struggles. I’m okay the way I am–I have an algorithm to follow.

Why write this post?

It would be an awfully surprising coincidence if mine was the *only* brain that worked this way. I’m not a special snowflake. And other people who interact with the Less Wrong community might not deal with it the way I do. They might try to twist their brains into the ‘right’ shape, and break their motivational system. Or they might decide that rationality is stupid and walk away.

False Friends and Tone Policing

TL;DR: *It can be helpful to reframe arguments about tone, trigger warnings, and political correctness as concerns about false cognates/false friends. You may be saying something that sounds innocuous to you, but translates to something much stronger/more vicious to your audience. Cultivating a debating demeanor that invites requests for tone concerns can give you more information about about the best way to avoid distractions and have a productive dispute.*

When I went on a two-week exchange trip to China, it was clear the cultural briefing was informed by whatever mistakes or misunderstandings had occurred on previous trips, recorded and relayed to us so that we wouldn't think, for example, that our host siblings were hitting on us if they took our hands while we were walking.

But the most memorable warning had to do with Mandarin filler words. While English speakers cover gaps with "uh" "um" "ah" and so forth, the equivalent filler words in Mandarin had an African-American student on a previous trip pulling aside our tour leader and saying he felt a little uncomfortable since his host family appeared to be peppering all of their comments with "nigga, nigga, nigga..."

As a result, we all got warned ahead of time. The filler word (那个 - nèige) was a false cognate that, although innocuous to the speaker, sounded quite off-putting to us. It helped to be warned, but it still required some deliberate, cognitive effort to remind myself that I wasn't actually hearing something awful and to rephrase it in my head.

When I've wound up in arguments about tone, trigger warnings, and taboo words, I'm often reminded of that experience in China. Limiting language can prompt suspicion of closing off conversations, but in a number of cases, when my friends have asked me to rephrase, it's because the word or image I was using was as distracting (however well meant) as 那个 was in Beijing.

It's possible to continue a conversation with someone who's every statement is laced with "nigga" but it takes effort. And no one is obligated to expend their energy on having a conversation with me if I'm making it painful or difficult for them, even if it's as the result of a false cognate (or, as the French would say, false friend) that sounds innocuous to me but *awful* to my interlocutor. If I want to have a debate at all, I need to stop doing the verbal equivalent of assaulting my friend to make any progress.

It can be worth it to pause and reconsider your language even if the offensiveness of a word or idea is exactly the subject of your dispute. When I hosted a debate on "R: Fire Eich" one of the early speakers made it clear that, in his opinion, opposing gay marriage was logically equivalent to endorsing gay genocide (he invoked a slippery slope argument back to the dark days of criminal indifference to AIDS).

Pretty much no one in the room (whatever their stance on gay marriage) agreed with this equivalence, but we could all agree it was pretty lucky that this person had spoken early in the debate, so that we understood how he was hearing our speeches. If every time someone said "conscience objection," this speaker was appending "to enable genocide," the fervor and horror with which he questioned us made a lot more

sense, and didn't feel like personal viciousness. Knowing how high the stakes felt to him made it *easier* to have a useful conversation.

This is a large part of why I objected to PZ Myers's deliberate obtuseness during the brouhaha he sparked when he asked readers to steal him a consecrated Host from a Catholic church so that he could desecrate it. PZ ridiculed Catholics for getting upset that he was going to "hurt" a piece of bread, even though the Eucharist is a fairly obvious example of a false cognate that is heard/received differently by Catholics and atheists. (After all, if it wasn't holy to *someone*, he wouldn't be able to profane it). In PZ's incident, it was although we had informed our Chinese hosts about the 那个/nigga confusion, and they had started using it *more* boisterously, so that it would be clearer to us that *they* didn't find it offensive.

We were only able to defuse the awkwardness in China for two reasons.

1. The host family was so nice, aside from this one provocation, that the student noticed he was confused and sought advice.
2. There was someone on hand who understood both groups well enough to serve as an interpreter.

In an ordinary argument (especially one that takes place online) it's up to you to be visibly virtuous enough that, if you happen to be using a vicious false cognate, your interlocutor will find that *odd*, not of a piece with your other behavior.

That's one reason my debating friend *did* bother explaining explicitly the connection he saw between opposition to gay marriage and passive support of genocide -- he trusted us enough to think that we *wouldn't* endorse the implications of our arguments if he made them obvious. In the P.Z. dispute, when Catholic readers found him as the result of the stunt, they didn't have any such trust.

It's nice to work to cultivate that trust, and to be the kind of person your friends *do* approach with requests for trigger warnings and tone shifts. For one thing, I *don't* want to use emotionally intense false cognates and not know it, any more than I would want to be gesticulating hard enough to strike my friend in the face without noticing. For the most part, I prefer to excise the distraction, so it's easier for both of us to focus on the heart of the dispute, but, even if you think that the controversial term is essential to your point, it's helpful to know it causes your friend pain, so you have the opportunity to salve it some other way.

P.S. Arnold Kling's *The Three Languages of Politics* is a short read and a nice introduction to what political language you're using that sounds like horrible false cognates to people rooted in different ideologies.

P.P.S. I've cross-posted this on my usual blog, but am trying out cross-posting to Discussion sometimes.

What resources have increasing marginal utility?

Most resources you might think to amass have decreasing marginal utility: for example, a marginal extra $1,000 means much more to you if you have $0 than if you have $100,000. That means you can safely apply the 80-20 rule to most resources: you only need to get some of the resource to get most of the benefits of having it.

At the most recent CFAR workshop, Val dedicated a class to arguing that one resource in particular has **increasing** marginal utility, namely **attention**. Initially, efforts to free up your attention have little effect: the difference between juggling 10 things and 9 things is pretty small. But once you've freed up most of your attention, the effect is larger: the difference between juggling 2 things and 1 thing is huge. Val also argued that because of this funny property of attention, most people likely undervalue the value of freeing up attention by orders of magnitude.

During a conversation later in the workshop I suggested another resource that might have increasing marginal utility, namely **trust**. A society where people abide by contracts 80% of the time is not 80% as good as a society where people abide by contracts 100% of the time; most of the societal value of trust (e.g. decreasing transaction costs) doesn't seem to manifest until people are pretty close to 100% trustworthy. The analogous way to undervalue trust is to argue that e.g. cheating on your spouse is not so bad, because only one person gets hurt. But cheating on spouses in general undermines the trust that spouses should have in each other, and the cumulative impact of even 1% of spouses cheating on the institution of marriage as a whole could be quite negative. (Lots of things about the world make more sense from this perspective: for example, it seems like one of the main practical benefits of religion is that it fosters trust.)

What other resources have increasing marginal utility? How undervalued are they?

Against utility functions

I think we should stop talking about utility functions.

In the context of ethics for humans, anyway. In practice I find utility functions to be, at best, an occasionally useful metaphor for discussions about ethics but, at worst, an idea that some people start taking too seriously and which actively makes them worse at reasoning about ethics. To the extent that we care about causing people to become better at reasoning about ethics, it seems like we ought to be able to do better than this.

The funny part is that the failure mode I worry the most about is already an entrenched part of the Sequences: it's fake utility functions. The soft failure is people who think they know what their utility function is and say bizarre things about what this implies that they, or perhaps all people, ought to do. The hard failure is people who think they know what their utility function is and then do bizarre things. I hope the hard failure is not very common.

It seems worth reflecting on the fact that the point of the foundational LW material discussing utility functions was to make people better at reasoning about AI behavior and not about human behavior.

Mathematics as a lossy compression algorithm gone wild

This is yet another half-baked post from my old draft collection, but feel free to Crocker away.

There is an old adage from Eugene Wigner known as the "Unreasonable Effectiveness of Mathematics". Wikipedia:

> the mathematical structure of a physical theory often points the way to further advances in that theory and even to empirical predictions.

The way I interpret is that it is possible to find an algorithm to compress a set of data points in a way that is also good at predicting other data points, not yet observed. In yet other words, a good approximation is, for some reason, sometimes also a good extrapolation. The rest of this post elaborates on this anti-Platonic point of view.

Now, this point of view is not exactly how most people see math. They imagine it as some near-magical thing that transcends science and reality and, when discovered, learned and used properly, gives one limited powers of clairvoyance. While only the select few wizard have the power to discover new spells (they are known as scientists), the rank and file can still use some of the incantations to make otherwise impossible things to happen (they are known as engineers).

This metaphysical view is colorfully expressed by Stephen Hawking:

> What is it that breathes fire into the equations and makes a universe for them to describe? The usual approach of science of constructing a mathematical model cannot answer the questions of why there should be a universe for the model to describe. Why does the universe go to all the bother of existing?

Should one interpret this as if he presumes here that math, in the form of "the equations" comes first and only then there is a physical universe for math to describe, for some values of "first" and "then", anyway? Platonism seems to reach roughly the same conclusions:

Wikipedia defines platonism as

> the philosophy that affirms the existence of abstract objects, which are asserted to "exist" in a "third realm *distinct both from the sensible external world and from the internal world of consciousness, and is the opposite of nominalism*

In other words, math would have "existed" even if there were no humans around to discover it. In this sense, it is "real", as opposed to "imagined by humans". Wikipedia on mathematical realism:

> mathematical entities exist independently of the human mind. Thus humans do not invent mathematics, but rather discover it, and any other intelligent beings in the universe would presumably do the same. In this

> point of view, there is really one sort of mathematics that can be discovered: triangles, for example, are real entities, not the creations of the human mind.

Of course, the debate on whether mathematics is "invented" or "discovered" is very old. Eliezer-2008 chimes in in http://lesswrong.com/lw/mq/beautiful_math/:

> To say that human beings "invented numbers" - or invented the structure implicit in numbers - seems like claiming that Neil Armstrong hand-crafted the Moon. The universe existed before there were any sentient beings to observe it, which implies that physics preceded physicists.

and later:

> The amazing thing is that math is a game without a designer, and yet it is eminently playable.

In the above, I assume that what Eliezer means by physics is not the science of physics (a human endeavor), but the laws according to which our universe came into existence and evolved. These laws are not the universe itself (which would make the statement "physics preceded physicists" simply "the universe preceded physicists", a vacuous tautology), but some separate laws governing it, out there to be discovered. If only we knew them all, we could create a copy of the universe from scratch, if not "for real", then at least as a faithful model. This **universe-making recipe** is then what physics (the laws, not science) is.

And these laws apparently require mathematics to be properly expressed, so mathematics must "exist" in order for the laws of physics to exist.

Is this the only way to think of math? I don't think so. Let us suppose that the physical universe is the only "real" thing, none of those Platonic abstract objects. Let is further suppose that this universe is (somewhat) predictable. Now, what does it mean for the universe to be predictable to begin with? Predictable by whom or by what? Here is one approach to predictability, based on agency: a small part of the universe (you, the agent) can construct/contain a model of some larger part of the universe (say, the earth-sun system, including you) and optimize its own actions (to, say, wake up the next morning just as the sun rises).

Does waking up on time count as doing math? Certainly not by the conventional definition of math. Do migratory birds do math when they migrate thousands of miles twice a year, successfully predicting that there would be food sources and warm weather once they get to their destination? Certainly not by the conventional definition of math. Now, suppose a ship captain lays a course to follow the birds, using maps and tables and calculations? Does this count as doing math? Why, certainly the captain would say so, even if the math in question is relatively simple. Sometimes the inputs both the birds and the humans are using are the same: sun and star positions at various times of the day and night, the magnetic field direction, the shape of the terrain.

What is the difference between what the birds are doing and what humans are doing? Certainly both make predictions about the universe and act on them. Only birds do this instinctively and humans consciously, by "applying math". But this is a statement about the differences in cognition, not about some Platonic mathematical objects. One can even say that birds perform the relevant math instinctively. But this is a rather slippery slope. By this definition amoebas solve the diffusion equation when they move along the sugar gradient toward a food source. While this view has merits, the

mathematicians analyzing certain aspects of the Navier-Stokes equation might not take kindly being compared to a protozoa.

So, like JPEG is a lossy image compression algorithm of the part of the universe which creates an image on our retina when we look at a picture, the collection of the Newton's laws is a lossy compression algorithm which describes how a thrown rock falls to the ground, or how planets go around the Sun. in both cases we, a tiny part of the universe, are able to model and predict a much larger part, albeit with some loss of accuracy.

What would it mean then for a Universe to not "run on math"? In this approach it means that in such a universe no subsystem can contain a model, no matter how coarse, of a larger system. In other words, such a universe is completely unpredictable from the inside. Such a universe cannot contain agents, intelligence or even the simplest life forms.

Now, to the "gone wild" part of the title. This is where the traditional applied math, like counting sheep, or calculating how many cannons you can arm a ship with before it sinks, or how to predict/cause/exploit the stock market fluctuations, becomes "pure math", or math for math's sake, be it proving the Pythagorean theorem or solving a Millennium Prize problem. At this point the mathematician is no longer interested in modeling a larger part of the universe (except insofar as she predicts that it would be a fun thing to do for her, which is probably not very mathematical).

Now, there is at least one serious objection to this "math is jpg" epistemology. It goes as follows: "in any universe, no matter how convoluted, 1+1=2, so clearly mathematics transcends the specific structure of a single universe". I am skeptical of this logic, since to me 1,+,= and 2 are semi-intuitive models running in our minds, which evolved to model the universe we live in. I can certainly imagine a universe where none of these concepts would be useful in predicting anything, and so they would never evolve in the "mind" of whatever entity inhabits it. To me mathematical concepts are no more universal than moral concepts: sometimes they crystallize into useful models, and sometimes they do not. Like the human concept of honor would not be useful to spiders, the concept of numbers (which probably is useful to spiders) would not be useful in a universe where size is not a well-defined concept (like something based on a Conformal Field Theory).

So the "Unreasonable Effectiveness of Mathematics" is not at all unreasonable: it reflects the predictability of our universe. Nothing "breathes fire into the equations and makes a universe for them to describe", the equations are but one way a small part of the universe predicts the salient features of a larger part of it. Rather, an interesting question is what features of a predictable universe enable agents to appear in it, and how complex and powerful can these agents get.

Flowers for Algernon

Daniel Keyes, the author of the short story *Flowers for Algernon*, and a novel of the same title that is its expanded version, died three days ago.

Keyes wrote many other books in the last half-century, but none achieved nearly as much prominence as the original short story (published in 1959) or the novel (came out in 1966).

It's probable that many or even most regulars here at Less Wrong read *Flowers for Algernon*: it's a very famous SF story, it's about enhanced intelligence, and it's been a middle/high school literature class staple in the US. But most != all, and past experience showed me that assumptions of cultural affinity are very frequently wrong. So in case you haven't read the story, I'd like to invite you explicitly to do so. It's rather short, and available at this link:

Flowers for Algernon

(I was surprised to find out that the original story is not available on Amazon. The expanded novelization is. If you wonder which version is better to read, I have no advice to offer)

(I will edit this post in a week or so to remove the link to the story and this remark)