Best of LessWrong: April 2013

Best of LessWrong: April 2013

Explicit and tacit rationality

Like Eliezer, I "do my best thinking into a keyboard." It starts with a burning itch to figure something out. I collect ideas and arguments and evidence and sources. I arrange them, tweak them, criticize them. I explain it all in my own words so I can understand it better. By then it is nearly something that others would want to read, so I clean it up and publish, say, How to Beat Procrastination. I write essays in the original sense of the word: "attempts."

This time, I'm trying to figure out something we might call "tacit rationality" (c.f. tacit knowledge).

I tried and failed to write a *good* post about tacit rationality, so I wrote a *bad* post instead — one that is basically a patchwork of somewhat-related musings on explicit and tacit rationality. Therefore I'm posting this article to LW Discussion. I hope the ensuing discussion ends up leading somewhere with more clarity and usefulness.

Three methods for training rationality

Which of these three options do you think will train rationality (i.e. systematized winning, or "winning-rationality") most effectively?

1. Spend one year reading and re-reading *The Sequences*, studying the math and cognitive science of rationality, and discussing rationality online and at Less Wrong meetups.
2. Attend a CFAR workshop, then spend the next year practicing those skills and other rationality habits every week.
3. Run a startup or small business for one year.

Option 1 seems to be pretty effective at training people to talk intelligently *about* rationality (let's call that "talking-rationality"), and it seems to inoculate people against some common philosophical mistakes.

We don't yet have any examples of someone doing Option 2 (the first CFAR workshop was May 2012), but I'd expect Option 2 — if actually executed — to result in more winning-rationality than Option 1, and also a modicum of talking-rationality.

What about Option 3? Unlike Option 2 or especially Option 1, I'd expect it to train almost no ability to talk intelligently about rationality. But I *would* expect it to result in relatively good winning-rationality, due to its tight feedback loops.

Talking-rationality and winning-rationality can come apart

> I've come to believe... that the best way to succeed is to discover what you love and then find a way to offer it to others in the form of service, working hard, and also allowing the energy of the universe to lead you.
>
> Oprah Winfrey

Oprah isn't known for being a rational thinker. She is a known peddler of pseudoscience, and she attributes her success (in part) to allowing "the energy of the universe" to lead her.

Yet she must be doing *something* right. Oprah is a true rags-to-riches story. Born in Mississippi to an unwed teenage housemaid, she was so poor she wore dresses made of potato sacks. She was molested by a cousin, an uncle, and a family friend. She became pregnant at age 14.

But in high school she became an honors student, won oratory contests and a beauty pageant, and was hired by a local radio station to report the news. She became the youngest-ever news anchor at Nashville's WLAC-TV, then hosted several shows in Baltimore, then moved to Chicago and within months her own talk show shot from last place to first place in the ratings there. Shortly afterward her show went national. She also produced and starred in several TV shows, was nominated for an Oscar for her role in a Steven Spielberg movie, launched her own TV cable network and her own magazine (the "most successful startup ever in the [magazine] industry" according to *Fortune*), and became the world's first female black billionaire.

I'd like to suggest that Oprah's climb probably didn't come *merely* through inborn talent, hard work, and luck. To get from potato sack dresses to the Forbes billionaire list, Oprah had to make thousands of pretty good decisions. She had to make pretty accurate guesses about the likely consequences of various actions she could take. When she was wrong, she had to correct course fairly quickly. In short, she had to be fairly *rational*, at least in some domains of her life.

Similarly, I know plenty of business managers and entrepreneurs who have a steady track record of good decisions and wise judgments, and yet they are religious, or they commit basic errors in logic and probability when they talk about non-business subjects.

What's going on here? My guess is that successful entrepreneurs and business managers and other people must have pretty good *tacit rationality*, even if they aren't very proficient with the "rationality" concepts that Less Wrongers tend to discuss on a daily basis. Stated another way, successful businesspeople make fairly rational decisions and judgments, even though they may confabulate rather silly *explanations* for their success, and even though they don't understand the math or science of rationality well.

LWers can probably outperform Mark Zuckerberg on the CRT and the Berlin Numeracy Test, but Zuckerberg is laughing at them from atop a huge pile of utility.

Explicit and tacit rationality

Patri Friedman, in Self-Improvement or Shiny Distraction: Why Less Wrong is anti-Instrumental Rationality, reminded us that skill acquisition comes from deliberate practice, and reading LW is a "shiny distraction," not deliberate practice. He said a *real* rationality practice would look more like... well, what Patri describes is basically CFAR, though CFAR didn't exist at the time.

In response, and again long before CFAR existed, Anna Salamon wrote Goals for which Less Wrong does (and doesn't) help. Summary: Some domains provide rich, cheap feedback, so you don't need much LW-style rationality to become successful in those domains. But many of us have goals in domains that don't offer rapid feedback: e.g. whether to buy cryonics, which 40-year investments are safe, which metaethics to endorse. For this kind of thing you need LW-style rationality. (We could also state this as "Domains with rapid feedback train tacit rationality with respect to those domains, but for domains without rapid feedback you've got to do the best you can with LW-style "explicit rationality".)

The good news is that you should be able to combine explicit and tacit rationality. Explicit rationality can help you realize that you should force tight feedback loops into whichever domains you want to succeed in, so that you can have develop good intuitions about how to succeed in those domains. (See also: Lean Startup or Lean Nonprofit methods.)

Explicit rationality could also help you realize that the cognitive biases most-discussed in the literature aren't necessarily the ones you should focus on ameliorating, as Aaron Swartz wrote:

> Cognitive biases cause people to make choices that are *most obviously* irrational, but not *most importantly* irrational... Since cognitive biases are the primary focus of research into rationality, rationality tests mostly measure how good you are at avoiding them... LW readers tend to be fairly good at avoiding cognitive biases... But there a whole series of much more important irrationalities that LWers suffer from. (Let's call them "practical biases" as opposed to "cognitive biases," even though both are ultimately practical and cognitive.)
>
> ...Rationality, properly understood, is in fact a predictor of success. Perhaps if LWers used success as their metric (as opposed to getting better at avoiding obvious mistakes), they might focus on their most important irrationalities (instead of their most obvious ones), which would lead them to be more rational and more successful.

Final scattered thoughts

- If someone is consistently winning, and not just because they have tons of wealth or fame, then maybe you should conclude they have pretty good tacit rationality even if their explicit rationality is terrible.
- The positive effects of tight feedback loops might trump the effects of explicit rationality training.
- Still, I suspect explicit rationality *plus* tight feedback loops could lead to the best results of all.
- I really hope we can develop a real rationality dojo.
- If you're reading this post, you're probably spending too *much* time reading Less Wrong, and too *little* time hacking your motivation system, learning social skills, and learning how to inject tight feedback loops into everything you can.

Minor, perspective changing facts

There's a lot of background mess in our mental pictures of the world. We try and be accurate on important issues, but a whole lot of the less important stuff we pick up from the media, the movies, and random impressions. And once these impressions are in our mental pictures, they just don't go away - until we find a fact that causes us to say "huh", and reassess.

Here are three facts that have caused that "huh" in me, recently,
and completely rearranged minor parts of my mental map. I'm sharing them here, because that experience is a valuable one.

1. Think terrorist attack on Israel - did the phrase "suicide bombing" spring to mind? If so, you're so out of fashion: the last suicide bombing in Israel was in 2008 - a year where dedicated suicide bombers managed the feat of killing a grand total of 1 victim. Suicide bombings haven't happened in Israel for over half a decade.
2. Large scale plane crashes seem to happen all the time, all over the world. They must happen at least a few times a year, in every major country, right? Well, if I'm reading this page right, the last time there was an airline crash in the USA that killed more that 50 people was... in 2001 (2 months after 9/11). Nothing on that scale since then. And though there has been crashes on route to/from Spain and France since then, it seems that major air crashes in western countries is something that essentially never happens.
3. The major cost of a rocket isn't the fuel, as I'd always thought. It seems that the Falcon 9 rocket costs $54 million per launch, of which fuel is only $0.2 million (or, as I prefer to think of it - I could sell my house to get enough fuel to fly to space). In the difference between those two prices, lies the potential for private spaceflight to low-Earth orbit.

Privileging the Question

Related to: Privileging the Hypothesis

> Remember the exercises in critical reading you did in school, where you had to look at a piece of writing and step back and ask whether the author was telling the whole truth? If you really want to be a critical reader, it turns out you have to step back one step further, and ask not just whether the author is telling the truth, but *why he's writing about this subject at all.*

-- Paul Graham

> There's an old saying in the public opinion business: we can't tell people what to think, but we can tell them what to think about.

-- Doug Henwood

> Many philosophers—particularly amateur philosophers, and ancient philosophers—share a dangerous instinct: If you give them a question, they try to answer it.

-- Eliezer Yudkowsky

Here are some political questions that seem to commonly get discussed in US media: should gay marriage be legal? Should Congress pass stricter gun control laws? Should immigration policy be tightened or relaxed?

These are all examples of what I'll call **privileged questions** (if there's an existing term for this, let me know): questions that someone has unjustifiably brought to your attention in the same way that a privileged hypothesis unjustifiably gets brought to your attention. The questions above are probably not the most important questions we could be answering right now, even in politics (I'd guess that the economy is more important). Outside of politics, many LWers probably think "what can we do about existential risks?" is one of the most important questions to answer, or possibly "how do we optimize charity?"

Why has the media privileged these questions? I'd guess that the media is incentivized to ask whatever questions will get them the most views. That's a very different goal from asking the most important questions, and is one reason to stop paying attention to the media.

The problem with privileged questions is that you only have so much attention to spare. Attention paid to a question that has been privileged funges against attention you could be paying to better questions. Even worse, it may not feel from the inside like anything is wrong: you can apply all of the epistemic rationality in the world to answering a question like "should Congress pass stricter gun control laws?" and never once ask yourself where that question came from and whether there are better questions you could be answering instead.

I suspect this is a problem in academia too. Richard Hamming once gave a talk in which he related the following story:

> Over on the other side of the dining hall was a chemistry table. I had worked with one of the fellows, Dave McCall; furthermore he was courting our secretary at the

> time. I went over and said, "Do you mind if I join you?" They can't say no, so I started eating with them for a while. And I started asking, "What are the important problems of your field?" And after a week or so, "What important problems are you working on?" And after some more time I came in one day and said, "If what you are doing is not important, and if you don't think it is going to lead to something important, why are you at Bell Labs working on it?" I wasn't welcomed after that; I had to find somebody else to eat with!

Academics answer questions that have been privileged in various ways: perhaps the questions their advisor was interested in, or the questions they'll most easily be able to publish papers on. Neither of these are necessarily well-correlated with the most important questions.

So far I've found one tool that helps combat the worst privileged questions, which is to ask the following counter-question:

What do I plan on doing with an answer to this question?

With the worst privileged questions I frequently find that the answer is "nothing," sometimes with the follow-up answer "signaling?" That's a bad sign. (**Edit:** but "nothing" is different from "I'm just curious," say in the context of an interesting mathematical or scientific question that isn't motivated by a practical concern. Intellectual curiosity can be a useful heuristic.)

(I've also found the above counter-question generally useful for dealing with questions. For example, it's one way to notice when a question should be dissolved, and asked of someone else it's one way to help both of you clarify what they actually want to know.)

Fermi Estimates

Just before the Trinity test, Enrico Fermi decided he wanted a rough estimate of the blast's power before the diagnostic data came in. So he dropped some pieces of paper from his hand as the blast wave passed him, and used this to estimate that the blast was equivalent to 10 kilotons of TNT. His guess was remarkably accurate for having so little data: the true answer turned out to be 20 kilotons of TNT.

Fermi had a knack for making roughly-accurate estimates with very little data, and therefore such an estimate is known today as a Fermi estimate.

Why bother with Fermi estimates, if your estimates are likely to be off by a factor of 2 or even 10? Often, getting an estimate within a factor of 10 or 20 is enough to make a decision. So Fermi estimates can save you a lot of time, especially as you gain more practice at making them.

Estimation tips

These first two sections are adapted from *Guestimation 2.0*.

Dare to be imprecise. Round things off enough to do the calculations in your head. I call this the spherical cow principle, after a joke about how physicists oversimplify things to make calculations feasible:

> Milk production at a dairy farm was low, so the farmer asked a local university for help. A multidisciplinary team of professors was assembled, headed by a theoretical physicist. After two weeks of observation and analysis, the physicist told the farmer, "I have the solution, but it only works in the case of spherical cows in a vacuum."

By the spherical cow principle, there are 300 days in a year, people are six feet (or 2 meters) tall, the circumference of the Earth is 20,000 mi (or 40,000 km), and cows are spheres of meat and bone 4 feet (or 1 meter) in diameter.

Decompose the problem. Sometimes you can give an estimate in one step, within a factor of 10. (How much does a new compact car cost? $20,000.) But in most cases, you'll need to break the problem into several pieces, estimate each of them, and then recombine them. I'll give several examples below.

Estimate by bounding. Sometimes it is easier to give lower and upper bounds than to give a point estimate. How much time per day does the average 15-year-old watch TV? I don't spend any time with 15-year-olds, so I haven't a clue. It could be 30 minutes, or 3 hours, or 5 hours, but I'm pretty confident it's more than 2 minutes and less than 7 hours (400 minutes, by the spherical cow principle).

Can we convert those bounds into an estimate? You bet. But we don't do it by taking the *average*. That would give us (2 mins + 400 mins)/2 = 201 mins, which is within a factor of 2 from our upper bound, but a factor *100* greater than our lower bound. Since our goal is to estimate the answer within a factor of 10, we'll probably be way off.

Instead, we take the *geometric mean* — the square root of the product of our upper and lower bounds. But square roots often require a calculator, so instead we'll take the *approximate* geometric mean (AGM). To do that, we average the coefficients and exponents of our upper and lower bounds.

So what is the AGM of 2 and 400? Well, 2 is 2×10^0, and 400 is 4×10^2. The average of the coefficients (2 and 4) is 3; the average of the exponents (0 and 2) is 1. So, the AGM of 2 and 400 is 3×10^1, or 30. The precise geometric mean of 2 and 400 turns out to be 28.28. Not bad.

What if the sum of the exponents is an odd number? Then we round the resulting exponent down, and multiply the final answer by three. So suppose my lower and upper bounds for how much TV the average 15-year-old watches had been 20 mins and 400 mins. Now we calculate the AGM like this: 20 is 2×10^1, and 400 is still 4×10^2. The average of the coefficients (2 and 4) is 3; the average of the exponents (1 and 2) is 1.5. So we round the exponent down to 1, and we multiple the final result by three: $3(3\times10^1)$ = 90 mins. The precise geometric mean of 20 and 400 is 89.44. Again, not bad.

Sanity-check your answer. You should always sanity-check your final estimate by comparing it to some reasonable analogue. You'll see examples of this below.

Use Google as needed. You can often quickly find the exact quantity you're trying to estimate on Google, or at least some *piece* of the problem. In those cases, it's probably not worth trying to estimate it *without* Google.

Fermi estimation failure modes

Fermi estimates go wrong in one of three ways.

First, we might badly overestimate or underestimate a quantity. Decomposing the problem, estimating from bounds, and looking up particular pieces on Google should protect against this. Overestimates and underestimates for the different pieces of a problem should roughly cancel out, especially when there are many pieces.

Second, we might model the problem incorrectly. If you estimate teenage deaths per year on the assumption that most teenage deaths are from suicide, your estimate will probably be way off, because most teenage deaths are caused by accidents. To avoid this, try to decompose each Fermi problem by using a model you're fairly confident of, even if it means you need to use more pieces or give wider bounds when estimating each quantity.

Finally, we might choose a nonlinear problem. Normally, we assume that if one object can get some result, then two objects will get twice the result. Unfortunately, this doesn't hold true for nonlinear problems. If one motorcycle on a highway can transport a person at 60 miles per hour, then 30 motorcycles can transport 30 people at 60 miles per hour. However, 10^4 motorcycles cannot transport 10^4 people at 60 miles per hour, because there will be a huge traffic jam on the highway. This problem is difficult to avoid, but with practice you will get better at recognizing when you're facing a nonlinear problem.

Fermi practice

When getting started with Fermi practice, I recommend estimating quantities that you can easily look up later, so that you can see how accurate your Fermi estimates tend to be. Don't look up the answer before constructing your estimates, though! Alternatively, you might allow yourself to look up particular pieces of the problem — e.g. the number of Sikhs in the world, the formula for escape velocity, or the gross world product — but not the final quantity you're trying to estimate.

Most books about Fermi estimates are filled with examples done by Fermi estimate experts, and in many cases the estimates were probably adjusted after the author looked up the true answers. This post is different. My examples below are estimates I made *before* looking up the answer online, so you can get a realistic picture of how this works from someone who isn't "cheating." Also, there will be no selection effect: I'm going to do four Fermi estimates for this post, and I'm not going to throw out my estimates if they are way off. Finally, I'm not all that practiced doing "Fermis" myself, so you'll get to see what it's like for a relative newbie to go through the process. In short, I hope to give you a realistic picture of what it's like to do Fermi practice when you're just getting started.

Example 1: How many new passenger cars are sold each year in the USA?

The classic Fermi problem is "How many piano tuners are there in Chicago?" This kind of estimate is useful if you want to know the approximate size of the customer base for a new product you might develop, for example. But I'm not sure anyone knows how many piano tuners there *really* are in Chicago, so let's try a different one we probably *can* look up later: "How many new passenger cars are sold each year in the USA?"

As with all Fermi problems, there are many different models we could build. For example, we could estimate how many new cars a dealership sells per month, and then we could estimate how many dealerships there are in the USA. Or we could try to estimate the annual demand for new cars from the country's population. Or, if we happened to have read how many Toyota Corollas were sold last year, we could try to build our estimate from there.

The second model looks more robust to me than the first, since I know roughly how many Americans there are, but I have no idea how many new-car dealerships there are. Still, let's try it both ways. (I *don't* happen to know how many new Corollas were sold last year.)

Approach #1: Car dealerships

How many new cars does a dealership sell per month, on average? Oofta, I dunno. To support the dealership's existence, I assume it has to be at least 5. But it's probably not more than 50, since most dealerships are in small towns that don't get much action. To get my point estimate, I'll take the AGM of 5 and 50. 5 is 5×10^0, and 50 is

5×10^1. Our exponents sum to an odd number, so I'll round the exponent down to 0 and multiple the final answer by 3. So, my estimate of how many new cars a new-car dealership sells per month is $3(5\times10^0) = 15$.

Now, how many new-car dealerships are there in the USA? This could be tough. I know several towns of only 10,000 people that have 3 or more new-car dealerships. I don't recall towns much smaller than that having new-car dealerships, so let's exclude them. How many cities of 10,000 people or more are there in the USA? I have no idea. So let's decompose this problem a bit more.

How many *counties* are there in the USA? I remember seeing a map of counties colored by which national ancestry was dominant in that county. (Germany was the most common.) Thinking of that map, there were definitely more than 300 counties on it, and definitely less than 20,000. What's the AGM of 300 and 20,000? Well, 300 is 3×10^2, and 20,000 is 2×10^4. The average of coefficients 3 and 2 is 2.5, and the average of exponents 2 and 4 is 3. So the AGM of 300 and 20,000 is $2.5\times10^3 = 2500$.

Now, how many towns of 10,000 people or more are there per county? I'm pretty sure the average must be larger than 10 and smaller than 5000. The AGM of 10 and 5000 is 300. (I won't include this calculation in the text anymore; you know how to do it.)

Finally, how many car dealerships are there in cities of 10,000 or more people, on average? Most such towns are pretty small, and probably have 2-6 car dealerships. The largest cities will have many more: maybe 100-ish. So I'm pretty sure the average number of car dealerships in cities of 10,000 or more people must be between 2 and 30. The AGM of 2 and 30 is 7.5.

Now I just multiply my estimates:

[15 new cars sold per month per dealership] × [12 months per year] × [7.5 new-car dealerships per city of 10,000 or more people] × [300 cities of 10,000 or more people per county] × [2500 counties in the USA] = 1,012,500,000.

A sanity check immediately invalidates this answer. There's no way that 300 million American citizens buy a *billion* new cars per year. I suppose they *might* buy 100 million new cars per year, which would be within a factor of 10 of my estimate, but I doubt it.

As I suspected, my first approach was problematic. Let's try the second approach, starting from the population of the USA.

Approach #2: Population of the USA

There are about 300 million Americans. How many of them own a car? Maybe 1/3 of them, since children don't own cars, many people in cities don't own cars, and many households share a car or two between the adults in the household.

Of the 100 million people who own a car, how many of them bought a *new* car in the past 5 years? Probably less than half; most people buy used cars, right? So maybe 1/4 of car owners bought a new car in the past 5 years, which means 1 in 20 car owners bought a new car in the past *year*.

100 million / 20 = 5 million new cars sold each year in the USA. That doesn't seem crazy, though perhaps a bit low. I'll take this as my estimate.

Now is your last chance to try this one on your own; in the next paragraph I'll reveal the true answer.

…

…

…

Now, I Google new cars sold per year in the USA. Wikipedia is the first result, and it says "In the year 2009, about 5.5 million new passenger cars were sold in the United States according to the U.S. Department of Transportation."

Boo-yah!

Example 2: How many fatalities from passenger-jet crashes have there been in the past 20 years?

Again, there are multiple models I could build. I could try to estimate how many passenger-jet flights there are per year, and then try to estimate the frequency of crashes and the average number of fatalities per crash. Or I could just try to guess the total number of passenger-jet crashes around the world per year and go from there.

As far as I can tell, passenger-jet crashes (with fatalities) almost always make it on the

TV news and (more relevant to me) the front page of Google News. Exciting footage and multiple deaths will do that. So working just from memory, it feels to me like there are about 5 passenger-jet crashes (with fatalities) per year, so maybe there were about 100 passenger jet crashes with fatalities in the past 20 years.

Now, how many fatalities per crash? From memory, it seems like there are usually two kinds of crashes: ones where *everybody* dies (meaning: about 200 people?), and ones where only about 10 people die. I think the "everybody dead" crashes are less common, maybe 1/4 as common. So the average crash with fatalities should cause (200×1/4)+(10×3/4) = 50+7.5 = 60, by the spherical cow principle.

60 fatalities per crash × 100 crashes with fatalities over the past 20 years = 6000 passenger fatalities from passenger-jet crashes in the past 20 years.

Last chance to try this one on your own...

...

...

...

A Google search again brings me to Wikipedia, which reveals that an organization called ACRO records the number of airline fatalities each year. Unfortunately for my purposes, they include fatalities from cargo flights. After more Googling, I tracked down Boeing's "Statistical Summary of Commercial Jet Airplane Accidents, 1959-2011," but that report excludes jets lighter than 60,000 pounds, and excludes crashes caused by hijacking or terrorism.

It appears it would be a major research project to figure out the true answer to our question, but let's at least estimate it from the ACRO data. Luckily, ACRO has statistics on which percentage of accidents are from passenger and other kinds of flights, which I'll take as a proxy for which percentage of *fatalities* are from different kinds of flights. According to that page, 35.41% of accidents are from "regular schedule" flights, 7.75% of accidents are from "private" flights, 5.1% of accidents are from "charter" flights, and 4.02% of accidents are from "executive" flights. I think that captures what I had in mind as "passenger-jet flights." So we'll guess that 52.28% of fatalities are from "passenger-jet flights." I won't round this to 50% because we're not doing a Fermi estimate right now; we're trying to *check* a Fermi estimate.

According to ACRO's archives, there were 794 fatalities in 2012, 828 fatalities in 2011, and... well, from 1993-2012 there were a total of 28,021 fatalities. And 52.28% of that number is 14,649.

So my estimate of 6000 was off by less than a factor of 3!

Example 3: How much does the New York state government spends on K-12 education every year?

How might I estimate this? First I'll estimate the number of K-12 students in New York, and then I'll estimate how much this should cost.

How many people live in New York? I seem to recall that NYC's greater metropolitan area is about 20 million people. That's probably most of the state's population, so I'll guess the total is about 30 million.

How many of those 30 million people attend K-12 public schools? I can't remember what the United States' population pyramid looks like, but I'll guess that about 1/6 of Americans (and hopefully New Yorkers) attend K-12 at any given time. So that's 5 million kids in K-12 in New York. The number attending private schools probably isn't large enough to matter for factor-of-10 estimates.

How much does a year of K-12 education cost for one child? Well, I've heard teachers don't get paid much, so after benefits and taxes and so on I'm guessing a teacher costs about $70,000 per year. How big are class sizes these days, 30 kids? By the spherical cow principle, that's about $2,000 per child, per year on teachers' salaries. But there are lots of other expenses: buildings, transport, materials, support staff, etc. And maybe some money goes to private schools or other organizations. Rather than estimate all those things, I'm just going to guess that about $10,000 is spent per child, per year.

If that's right, then New York spends $50 billion per year on K-12 education.

Last chance to make your own estimate!

…

…

…

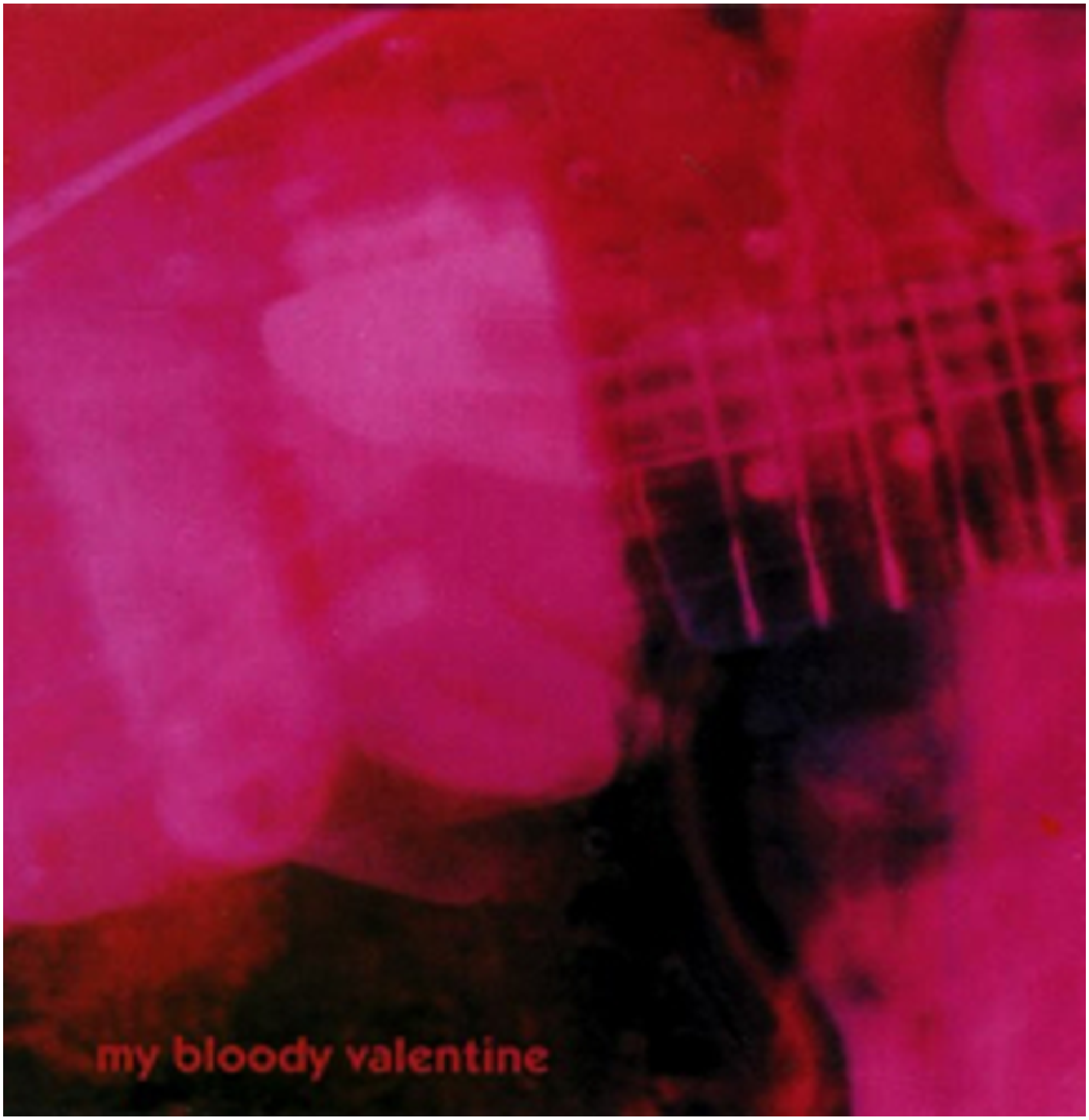
Before I did the Fermi estimate, I had Julia Galef check Google to find this statistic, but she didn't give me any hints about the number. Her two sources were Wolfram Alpha and a web chat with New York's Deputy Secretary for Education, both of which put the figure at approximately $53 billion.

Which is definitely within a factor of 10 from $50 billion. :)

Example 4: How many plays of My Bloody Valentine's "Only Shallow" have been reported to last.fm?

Last.fm makes a record of every audio track you play, if you enable the relevant feature or plugin for the music software on your phone, computer, or other device. Then, the service can show you charts and statistics about your listening patterns, and make personalized music recommendations from them. My own charts are here. (Chuck Wild / Liquid Mind dominates my charts because I used to listen to that artist while sleeping.)

My Fermi problem is: How many plays of "Only Shallow" have been reported to last.fm?



My Bloody Valentine is a popular "indie" rock band, and "Only Shallow" is probably one of their most popular tracks. How can I estimate how many plays it has gotten on last.fm?

What do I know that might help?

- I know last.fm is popular, but I don't have a sense of whether they have 1 million users, 10 million users, or 100 million users.
- I accidentally saw on Last.fm's Wikipedia page that just over 50 billion track plays have been recorded. We'll consider that to be one piece of data I looked up to help with my estimate.
- I seem to recall reading that major music services like iTunes and Spotify have about 10 million tracks. Since last.fm records songs that people play from their private collections, whether or not they exist in popular databases, I'd guess that

> the total number of different tracks named in last.fm's database is an order of magnitude larger, for about 100 million tracks named in its database.

I would guess that track plays obey a power law, with the most popular tracks getting vastly more plays than tracks of average popularity. I'd also guess that there are maybe 10,000 tracks more popular than "Only Shallow."

Next, I simulated being good at math by having Qiaochu Yuan show me how to do the calculation. I also allowed myself to use a calculator. Here's what we do:

Plays(rank) = C/(rank^P)

P is the exponent for the power law, and C is the proportionality constant. We'll guess that P is 1, a common power law exponent for empirical data. And we calculate C like so:

C ≈ [total plays]/ln(total songs) ≈ 2.5 billion

So now, assuming the song's rank is 10,000, we have:

Plays(10^4) = $2.5 \times 10^9/(10^4)$

Plays("Only Shallow") = 250,000

That seems high, but let's roll with it. Last chance to make your own estimate!

...

...

...

And when I check the answer, I see that "Only Shallow" has about 2 million plays on last.fm.

My answer was off by less than a factor of 10, which for a Fermi estimate is called *victory*!

Unfortunately, last.fm doesn't publish all-time track rankings or other data that might help me to determine which parts of my model were correct and incorrect.

Further examples

I focused on examples that are similar in structure to the kinds of quantities that entrepreneurs and CEOs might want to estimate, but of course there are all kinds of things one can estimate this way. Here's a sampling of Fermi problems featured in various books and websites on the subject:

Play Fermi Questions: 2100 Fermi problems and counting.

Guesstimation (2008): If all the humans in the world were crammed together, how much area would we require? What would be the mass of all 10^8 MongaMillions lottery tickets? On average, how many people are airborne over the US at any given moment? How many cells are there in the human body? How many people in the world are picking their nose right now? What are the relative costs of fuel for NYC rickshaws and automobiles?

*Guesstimation 2.0 (2011): If we launched a trillion one-dollar bills into the atmosphere, what fraction of sunlight hitting the Earth could we block with those dollar bills? If a million monkeys typed randomly on a million typewriters for a year, what is the longest string of consecutive correct letters of *The Cat in the Hat* (starting from the beginning) would they likely type? How much energy does it take to crack a nut? If an airline asked its passengers to urinate before boarding the airplane, how much fuel would the airline save per flight? What is the radius of the largest rocky sphere from which we can reach escape velocity by jumping?

How Many Licks? (2009): What fraction of Earth's volume would a mole of hot, sticky, chocolate-jelly doughnuts be? How many miles does a person walk in a lifetime? How many times can you outline the continental US in shoelaces? How long would it take to read every book in the library? How long can you shower and still make it more environmentally friendly than taking a bath?

Ballparking (2012): How many bolts are in the floor of the Boston Garden basketball court? How many lanes would you need for the outermost lane of a running track to be the length of a marathon? How hard would you have to hit a baseball for it to never land?

University of Maryland Fermi Problems Site: How many sheets of letter-sized paper are used by all students at the University of Maryland in one semester? How many blades of grass are in the lawn of a typical suburban house in the summer? How many golf balls can be fit into a typical suitcase?

Stupid Calculations: a blog of silly-topic Fermi estimates.

Conclusion

Fermi estimates can help you become more efficient in your day-to-day life, and give you increased confidence in the decisions you face. If you want to become proficient in making Fermi estimates, I recommend practicing them 30 minutes per day for three months. In that time, you should be able to make about (2 Fermis per day)×(90 days) = 180 Fermi estimates.

If you'd like to write down your estimation attempts and then publish them here, please do so as a reply to this comment. One Fermi estimate per comment, please!

Alternatively, post your Fermi estimates to the dedicated subreddit.

Update 03/06/2017: I keep getting requests from professors to use this in their classes, so: I license anyone to use this article noncommercially, so long as its authorship is noted (me = Luke Muehlhauser).

Litany of a Bright Dilettante

So, one more litany, hopefully someone else finds it as useful.

It's an understatement that humility is not a common virtue in online discussions, even, or especially when it's most needed.

I'll start with my own recent example. I thought up a clear and obvious objection to one of the assertions in Eliezer's critique of the FAI effort compared with the Pascal's Wager and started writing a witty reply. ...And then I stopped. In large part because I had just gone through the same situation, but on the other side, dealing with some of the comments to my post about time-turners and General Relativity by those who know next to nothing about General Relativity. It was irritating, yet here I was, falling into the same trap. And not for the first time, far from it. The following is the resulting thought process, distilled to one paragraph.

I have not spent 10,000+ hours thinking about this topic in a professional, all-out, do-the-impossible way. I probably have not spent even one hour seriously thinking about it. I probably do not have the prerequisites required to do so. I probably don't even know what prerequisites are required to think about this topic productively. In short, there are almost guaranteed to exist unknown unknowns which are bound to trip up a novice like me. The odds that I find a clever argument contradicting someone who works on this topic for a living, just by reading one or two popular explanations of it are minuscule. So if I think up such an argument, the odds of it being both new and correct are heavily stacked against me. It is true that they are non-zero, and there are popular examples of non-experts finding flaws in an established theory where there is a consensus among the experts. Some of them might even be true stories. No, Einstein was not one of these non-experts, and even if he were, I am not Einstein.

And so on. So I came up with the following, rather unpolished mantra:

If I think up what seems like an obvious objection, I will resist assuming that I have found a Weaksauce Weakness in the experts' logic. Instead I may ask politely whether my argument is a valid one, and if not, where the flaw lies.

If you think it useful, feel free to improve the wording.

Problems in Education

Post will be returning in Main, after a rewrite by the company's writing staff. Citations Galore.

We Don't Have a Utility Function

Related: Pinpointing Utility

If I ever say "my utility function", you could reasonably accuse me of cargo-cult rationality; trying to become more rational by superficially immitating the abstract rationalists we study makes about as much sense as building an air traffic control station out of grass to summon cargo planes.

There are two ways an agent could be said to have a utility function:

1. It could behave in accordance with the VNM axioms; always choosing in a sane and consistent manner, such that "there exists a U". The agent need not have an explicit representation of U.

2. It could have an explicit utility function that it tries to expected-maximize. The agent need not perfectly follow the VNM axioms all the time. (Real bounded decision systems will take shortcuts for efficiency and may not achieve perfect rationality, like how real floating point arithmetic isn't associative).

Neither of these is true of humans. Our behaviour and preferences are not consistent and sane enough to be VNM, and we are generally quite confused about what we even want, never mind having reduced it to a utility function. Nevertheless, you still see the occasional reference to "my utility function".

Sometimes "my" refers to "abstract me who has solved moral philosophy and or become perfectly rational", which at least doesn't run afoul of the math, but is probably still wrong about the particulars of what such an abstract idealized self would actually want. But other times it's a more glaring error like using "utility function" as shorthand for "entire self-reflective moral system", which may not even be VNMish.

But this post isn't really about all the ways people misuse terminology, it's about where we're actually at on the whole problem for which a utility function might be the solution.

As above, I don't think any of us have a utility function in either sense; we are not VNM, and we haven't worked out what we want enough to make a convincing attempt at trying. Maybe someone out there has a utility function in the second sense, but I doubt that it actually represents what they would want.

Perhaps then we should speak of what we want in terms of "terminal values"? For example, I might say that it is a terminal value of mine that I should not murder, or that freedom from authority is good.

But what does "terminal value" mean? Usually, it means that the value of something is not contingent on or derived from other facts or situations, like for example, I may value beautiful things in a way that is not derived from what they get me. The recursive chain of valuableness *terminates* at some set of values.

There's another connotation, though, which is that your terminal values are akin to *axioms*; not subject to argument or evidence or derivation, and simply given, that there's no point in trying to reconcile them with people who don't share them. This is the meaning people are sometimes getting at when they explain failure to agree with someone as "terminal value differences" or "different set of moral axioms". This is completely reasonable, if and only if that is in fact the nature of the beliefs in question.

About two years ago, it very much felt like freedom from authority was a terminal value for me. Those hated authoritarians and fascists were simply *wrong*, probably due to some fundamental neurological fault that could not be reasoned with. The very prototype of "terminal value differences".

And yet here I am today, having been reasoned out of that "terminal value", such that I even appreciate a certain aesthetic in bowing to a strong leader.

If that was a terminal value, I'm afraid the term has lost much of its meaning to me. If it was not, if even the most fundamental-seeming moral feelings are subject to argument, I wonder if there is any coherent sense in which I could be said to have terminal values at all.

The situation here with "terminal values" is a lot like the situation with "beliefs" in other circles. Ask someone what they believe in most confidently, and they will take the opportunity to differentiate themselves from the opposing tribe on uncertain controversial issues; god exists, god does not exist, racial traits are genetic, race is a social construct. The pedant answer of course is that the sky is probably blue, and that that box over there is about a meter long.

Likewise, ask someone for their terminal values, and they will take the opportunity to declare that those hated greens are utterly wrong on morality, and blueness is wired into their very core, rather than the obvious things like beauty and friendship being valuable, and paperclips not.

So besides not having a utility function, those aren't your terminal values. I'd be suprised if even the most pedantic answer weren't subject to argument; I don't seem to have anything like a stable and non-negotiable value system at all, and I don't think that I am even especially confused relative to the rest of you.

Instead of a nice consistent value system, we have a mess of intuitions and hueristics and beliefs that often contradict, fail to give an answer, and change with time and mood and memes. And that's all we have. One of the intuitions is that we want to fix this mess.

People have tried to do this "Moral Philosophy" thing before, myself included, but it hasn't generally turned out well. We've made all kinds of overconfident leaps to what turn out to be unjustified conclusions (utilitarianism, egoism, hedonism, etc), or just ended up wallowing in confused despair.

The zeroth step in solving a problem is to notice that we have a problem.

The problem here, in my humble opinion, is that we have no idea what we are doing when we try to do Moral Philosophy. We need to go up a meta-level and get a handle on Moral MetaPhilosophy. What's the problem? What are the relevent knowns? What are the unknowns? What's the solution process?

Ideally, we could do for Moral Philosphy approximately what Bayesian probability theory has done for Epistemology. My moral intuitions are a horrible mess, but so are my epistemic intuitions, and yet we more-or-less know what we are doing in epistemology. A problem like this has been solved before, and this one seems solvable too, if a bit harder.

It might be that when we figure this problem out to the point where we can be said to have a consistent moral system with real terminal values, we will end up with a utility function, but on the other hand, we might not. Either way, let's keep in mind that we are still on rather shaky ground, and at least refrain from believing the confident declarations of moral wisdom that we so like to make.

Moral Philosophy is an important problem, but the way is not clear yet.

Problems in Education

Alright guys. The main complaint of the discussion article was simply "hoax", yelled as loudly or as quietly as the user felt about it. Hopefully this won't get the same treatment.

We have been evaluating educational, grant-funded programs for 20 years. Throughout these years, we have witnessed a slow change in how students are selected for academic services. Traditionally, students were targeted for academic services and opportunities based on demographic characteristics—usually race and, until recently, family income status (based on free or reduced priced lunch). Wealthier, white students are given challenging lessons and tracked into the advanced courses, while their non-white and poorer peers are tracked low and given remediation services. The latter students are often referred to as “at-risk,” though we are finding more and more that the greatest risk these students face is being placed into inappropriate remedial courses which eventually bar them from access to advanced courses. After students have been labeled “at-risk,” and then tracked inappropriately and provided unnecessary (and often harmful) remediation, their downward trajectory continues throughout their education. The demographic gap this creates continues to expand, despite the lip service and excessive tax and grant funds paid to eliminate—or at least lessen—this very gap. This “at-risk” model of assigning services is slowly being replaced by a “pro-equity” model. The driving force behind this change is the availability and use of data.

The literature is full of documentation that certain demographic groups have traditionally had less access to advanced math and science courses than equally scoring students belonging to demographic groups thought to be “not at risk.” Some examples from research follow.

- Sixth grade course placement is the main predictor of eighth grade course placement, and social factors--mainly race---are key predictors of sixth grade course placement (O’Connor, Lewis, & Mueller, 2007).
- Among low-income students, little is done to assess which are high achievers. Few programs are aimed at them, and their numbers are lumped in with “adequate” achievers in No Child Left Behind reporting. As a result, little is known about effective practices for low-income students (Wyner, Bridgeland, & DiIulio Jr., 2007).
- In a California school district, researchers found that of students who demonstrated the ability to be admitted to algebra, 100% of the Asians, 88% of the whites, 51% of the Blacks, and 42% of the Latinos were admitted (Stone & Turba, 1999).
- Tracking has been described as “a backdoor device for sorting students by race and class.” Many researchers agree (Abu El-Haj & Rubin, 2009).
- When course grades are used to determine placement, studies show that some students’ grades “matter” more than others. Perceptions of race and social class are often used to determine placement (Mayer, 2008).
- Studies show that when schools allow students the freedom to choose which track they’ll take, teachers and counselors discourage many previously lower tracked students from choosing the higher track (Yonezawa, Wells, & Serna, 2002).
- The sequence of math students take in middle school essentially determines their math track for high school. In North Carolina, this is true because of math prerequisites for higher level math (North Carolina Department of Public Instruction, 2009).

We are seeing a move toward using objective data for placement into gateway courses, such as 8th grade algebra. Many school districts are beginning to use Education Value Added Assessment (EVAAS) and other data system scores that predict success in 8th grade algebra for criteria to enroll. This pro-equity model is replacing the traditional, at-risk model that relied on professional judgment. One example of this is in Wake County, North Carolina. Superintendent Tony Tata attributed a 44% increase in the number of students enrolled in algebra to the use of the predictive software, EVAAS, to identify students likely to be successful. The success rate in the course increased with the addition of these students (KeungHu, 2012).

Although the pro-equity model of using objective data to assign students to more rigorous courses has proven successful, many people resist it. These people cling to the at-risk model, dismissing the objective data as inconclusive. Many of the overlooked students who were predicted to succeed, yet were placed in lower tracks (disproportionately minorities), are “weaker,” according to the old-school staff, and allowing these students into the gateway 8th-grade algebra course would be a disservice to them. (Not allowing them into this course ensures their bleak academic future.) Review of the data had shown that strong students were being overlooked, and this objective use of data helps identify them (Sanders, Rivers, Enck, Leandro, & White, 2009).

The changes in education began with concern for aligning academic services with academic need. Aligning opportunities for rigor and enrichment is only just beginning. In the past, a large proportion of federal grant funds were for raising proficiency rates. In the at-risk model, grant funds were provided for services to the minority and poor demographic groups with the goals of raising academic proficiency rates. When we first started evaluating grant-funded programs, most federal grants were entirely in the at-risk model. The students were targeted for services based on demographic characteristics. The goals were to deliver the services to this group. Staff development was often designed to help staff understand children in poverty and what their lives are like, rather than helping them learn how to deliver an effective reading or math intervention. The accountability reports we were hired to write consisted of documentation that the correct demographic group was served, the program was delivered, and staff received their professional development. Proficiency rates were rarely a concern.

In 2004, the federal government developed the Program Assessment Rating Tool (PART) to provide accountability to grant-funded programs by rating their effectiveness. The PART system assigned scores to programs based on services being related to goals, showing that the goals were appropriate for the individuals served, and student success measured against quality standards and assessments. PART rated programs that could not demonstrate whether they have been effective or not because of lack of data or clear performance goals with the rating “Results Not Demonstrated” (U.S. Office of Management and Budget and Federal Agencies, n.d. "The Program Assessment Rating Tool") . In 2009, nearly half (47%) of U.S. Department of Education grant programs rated by the government are given this rating, thus illustrating the difficulties of making this transition to outcome based accountability (U.S. Office of Management and Budget and Federal agencies, n.d. "Department of Education programs"). The earliest changes were in accountability, not in program services or how to target students. Accountability reports began asking for pre- and post-comparisons of academic scores. For example, if funds were for raising the proficiency rates in reading, then evaluation reports were required to compare pre- and post-reading scores. This was a confusing period, because programs still targeted students based on demographic information and provided services that often had no research basis linking them to academic achievement; professional development often remained focused on empathizing with children in poverty, although the goals and objectives would now be written in terms of the participants raising their academic achievement to proficiency. We evaluators were often called in at the conclusion of programs to compare pre- and post-academic scores, and determine whether participants improved their scores to grade-level proficiency. We often saw the results of capable students treated like low-achievers, thought to have no self-esteem, and provided remedial work. Such treatment damaged the participants who had previously scored at or above proficient prior to services.

A typical narrative of an evaluation might read:

The goal of the program was to raise the percentage of students scoring proficient in reading. The program targeted and served low-income and minority students. Staff received professional development on understanding poor children. Services offered to students included remedial tutorials and esteem-building activities. When the program ended, pre-reading scores were obtained and compared with post-scores to measure progress toward the program objective. At that time, it was discovered that a large percentage of participants were proficient prior to receiving services.

Rather than cite our own evaluations, we found many examples from the school districts reporting on themselves.

Accelerated Learning Program.

The following is a direct quote from a school system in North Carolina:

. . . Although ALP [Accelerated Learning Program] was designed primarily to help students reach proficiency as measured by End-of-Grade (EOG) tests, only 41.1% of those served showed below-grade-level scores on standard tests before service in literacy. In mathematics, 73.3% of students served had below-grade-level scores. ALP served about 40% of students who scored below grade level within literacy and within mathematics, with other services supporting many others. . . . Compared to those not served, results for Level I-II students were similar, but results for Level III-IV students were less positive. One third of non- proficient ALP mathematics students reached proficiency in 2008, compared to 42.1% of other students. (Lougee & Baenen, 2009).

Foundations of Algebra

This program was designed for students who fit specific criteria, yet it served many students who did not. Students who were below proficient or almost proficient were to be placed in courses to eventually prepare them for Algebra I. When criteria for placement are not met, determining program effectiveness is difficult, if not impossible. Students were likely entered into the program based on teacher recommendations, which were subsequently based on demographic factors such as race. The teachers “mistook” these students for below-proficient students when they were not. Had objective data, such as actual proficiency scores, been consulted, the proper students could have been served. The report indicates a success, as a higher percentage of these students than similar students who were not served enrolled in Algebra I. However, it is not known if this comparison group includes only students who actually meet the criteria, or if they are a heterogeneous mix of students of varying abilities. Missing data also makes program effectiveness evaluation difficult (Paeplow, 2010).

Partnership for Educational Success (PES)

This program was purportedly for students who are “at risk,” which is defined as students who scored below grade level on EOG (below proficiency) and have been “identified by the PES team as having family issues that interfere with school success.” What is meant by “family issues” is unclear. The majority of students served are Economically Disadvantaged (ED) (91.3%) and Black (71.5%). More than half the students served, according to the evaluation, were at or above grade level on their EOGs when they began the program, thus making program effectiveness difficult to judge. The family component is an integral part of the program, and outside agencies visit families. Many community organizations are involved. But if the staff could miss so easy a datum as EOG scores for so many students, one has to wonder about such a subjective criterion as “family issues.” The program appears to have targeted ED students, with little regard to prior performance data. Data for many students (43.5%) was missing. Teachers indicate that parents of the targeted families have become more involved in the school, but little else has changed (Harlow & Baenen, 2004).

Helping Hands

Helping Hands was initiated based on data indicating that Black males lag behind other groups in academic achievement. The program is supposed to serve Black males, and most of the participants fit these criteria. The program is also designed to improve academics, and to curtail absenteeism and suspensions. Although the percentage of selected participants who needed improvement in these areas was higher than it was for the overall population of the students served, not all students served demonstrated a need for intervention. Many students were at grade level, were not chronically absent, and had not been suspended. Yet they were served because they were Black and male (Paeplow, 2009).

At Hodge Road Elementary School, students were tutored with remedial work in an after-school program. The only criterion the students had to meet to be allowed into the program was the inability to pay full price for their lunch. Their academic performance was irrelevant. (To be fair, these criteria were instituted by No Child Left Behind, and not the school system.) Most students were already reading and doing math at or above grade level (the two subjects for which tutoring was provided). The evaluation shows that giving remedial coursework to students who are at or above grade level, as if they were below grade level, can actually harm them. In the final statistics, 11.1% of Level III & IV 3rd through 5th graders scored below grade level after being served, compared with only 2% of a comparable group who were not served. An astonishing 23% of students in kindergarten through 2nd grade served who were at or above grade level prior to the tutoring scored below grade level afterward, compared with 8% of comparable students who were not served (Paeplow & Baenen, 2006).

AVID

AVID is a program designed for students who may be the first in their families to attend college, and who are average academic performers. The program, developed in the 1980s, maintains that by providing support while holding students to high academic standards, the achievement gap will narrow as students succeed academically and go on to successfully complete higher level education. Fidelity of implementation is often violated, which, as proponents admit on AVID’s own website (www.AVID.org) may compromise the entire program. Student participants must have a GPA of 2.0-3.5. We were asked to evaluate Wake County Public School Systems AVID program. Many students chosen for the program, however, did not fit the criteria (Lougee & Baenen, 2008). Because AVID requirements were not met, a meaningful evaluation was not possible.

This AVID program was implemented with the goal of increasing the number of under-represented students in 8th grade algebra. This was at a time when no criteria for enrollment in 8th grade algebra existed (i.e., a target to help the students reach didn't exist), and high scoring students in this very group were not being referred for enrollment in algebra. Under these conditions, the program makes no sense. In summary, the goal of this program is to enroll in 8th grade algebra more low-income, minority, and students whose parents didn't go to college. Only students recommended by teachers can enroll in 8th grade algebra. The data showed that very high-scoring, low-income and minority students were not being recommended for 8th grade algebra. Why do we think that students whose parents didn't go to college can't enroll in 8th grade algebra without being in an intervention program first? (Also, how it is determined that the students' parents did not attend college is not addressed.) The program is for low-average students. They served high-average students. Then they still didn't recommend them to be in 8th grade algebra. This program is very expensive. We have evaluated this program in many school districts and we find the same results, typically, as this report.

During this era, the interventions typically have not been related to the desired outcomes by research. For example, self-esteem-building activities were often provided to increase the odds of passing a math class, or to improve reading scores. Sometimes programs would be academic, but claims for success were not research-based, nor was the relationship between the activities and the desired outcomes. Although many interventions were at least related to the academic subject area the program was trying to impact, it was not unheard of to see relaxation courses alone for increasing math test scores, or make-overs and glamor shots for raising self-esteem, which in turn would allegedly raise reading scores.

During the last decade, education has slowly moved toward requiring accountability in terms of comparing pre- and post-scores. We saw this causing confusion and fear, rather than clarity. More than once, when we reported to school districts that they had served significant numbers of students who were already at or above proficiency levels, they thought we were saying they had served high-income students instead of their target population of low-income students. We have seen many school systems assess their own programs, write evaluation reports like the examples above, and then continue to implement the programs without any changes. We have worked with some educators whose eyes were opened to the misalignment of services and needs, and they learned to use data, to identify appropriate interventions, and keep records to make accountability possible. We've seen these innovators close their achievement gaps while raising achievement of the top. But, those around them didn't see this as replicable.

Race to the Top will impact the rate of change from the at-risk to the pro-equity model. Teacher and principal evaluations are going to include measures of growth in student learning (White House Office of the Press Secretary, 2009). EVAAS will be used to measure predicted scores with observed scores. If high-achieving students who are predicted to succeed in 8th grade algebra are tracked into the less rigorous 9th grade algebra, they are not likely to make their predicted growth .

We are moving out of this era, and the pace of change toward identifying student needs using appropriate data is picking up. North Carolina's newly legislated program, Read to Achieve, mandates that reading interventions for students in K-3 be aligned to the literacy skills the students struggle with, and that data be used to determine whether students are struggling with literacy skills. Schools must also keep records for accountability. Although this approach seems logical, it is quite innovative compared with the past reading interventions that targeted the wrong students (North Carolina State Board of Education; Department of Public Instruction, n.d.).

Education Grant programs are now requiring that applicants specify what data they will use to identify their target population, and how the intervention relates to helping the participants achieve the program goals. Staff development must relate to delivering the services well, and accountability must show that these things all happened correctly, while documenting progress toward the program objectives. It is a new era. We are not there yet, but it is coming.

References

Harlow, K., & Baenen, N. (2004). E & R Report No. 04.09: Partnership for Educational Success 2002-03: Implementation and outcomes. Raleigh, NC: Wake County Public School System. Retrieved from http://www.wcpss.net/evaluation-research/reports/2004/0409partnership_edu.pdf

KeungHu. (2012). Wake County Superintendent Tony Tata on gains in Algebra I enrollment and proficiency. Retrieved from http://blogs.newsobserver.com/wakeed/wake-county-superintendent-tony-tata-on-gains-in-algebra-i-enrollment-and-proficiency

Lougee, A., & Baenen, N. (2008). E & R Report No. 08.07: Advancement Via Individual Determination (AVID): WCPSS Program Evaluation. Retrieved from http://www.wcpss.net/evaluation-research/reports/2008/0807avid.pdf

Lougee, A., & Baenen, N. (2009). E&R Report No. 09.27: Accelerated Learning Program (ALP) grades 3-5: Evaluation 2007-08. Retrieved from http://www.wcpss.net/evaluation-research/reports/2009/0927alp3-5_2008.pdf

Mayer, A. (2008). Understanding how U.S. secondary schools sort students for instructional purposes: Are all students being served equally? . American Secondary Education , 36(2), 7–25.

North Carolina Department of Public Instruction. (2009). Course and credit requirements. Retrieved from http://www.ncpublicschools.org/curriculum/graduation

North Carolina State Board of Education; Department of Public Instruction. (n.d.). North Carolina Read to Achieve: A guide to implementing House Bill 950/S.L. 2012-142 Section 7A. Retrieved from https://eboard.eboardsolutions.com/Meetings/Attachment.aspx?S=10399&AID=11774&MID=783

O’Connor, C., Lewis, A., & Mueller, J. (2007). Researching “Black” educational experiences and outcomes: Theoretical and methodological considerations. Educational Researcher. Retrieved from http://www.sociology.emory.edu/downloads/O%5c’Connor_Lewis_Mueller_2007_Researching_black_educational_experiences_and_outcomes_theoretical_and_methodological_considerations.pdf

Paeplow, C. (2009). E & R Report No. 09.30: Intervention months grades 6-8: Elective results 2008-09. Raleigh, NC: Wake County Public School System. Retrieved from http://www.wcpss.net/evaluation-research/reports/2009/0930imonths6-8.pdf

Paeplow, C. (2010). E & R Report No. 10.28: Foundations of Algebra: 2009-10. Raleigh, NC: Wake County Public School System. Retrieved from http://assignment.wcpss.net/results/reports/2011/1028foa2010.pdf

Paeplow, C., & Baenen, N. (2006). E & R Report No. 06.09: Evaluation of Supplemental Educational Services at Hodge Road Elementary School 2005-06. Raleigh. Retrieved from http://www.wcpss.net/evaluation-research/reports/2006/0609ses_hodge.pdf

Sanders, W. L., Rivers, J. C., Enck, S., Leandro, J. G., & White, J. (2009). Educational Policy Brief: SAS® Response to the “WCPSS E & R Comparison of SAS © EVAAS © Results and WCPSS Effectiveness Index Results,” Research Watch, E&R Report No. 09.11, March 2009. Cary, NC: SAS. Retrieved from http://content.news14.com/pdf/sas_report.pdf

Stone, C. B., & Turba, R. (1999). School counselors using technology for advocacy. Journal of Technology in Counseling. Retrieved from http://jtc.colstate.edu/vol1_1/advocacy.htm

U.S. Office of Management and Budget and Federal Agencies. (n.d.). The Program Assessment Rating Tool (PART). Retrieved from http://www.whitehouse.gov/omb/expectmore/part.html

U.S. Office of Management and Budget and Federal agencies. (n.d.). Department of Education programs. Retrieved from http://www.whitehouse.gov/omb/expectmore/agency/018.html

White House Office of the Press Secretary. (2009). Fact Sheet: The Race to the Top. Washington D.C. Retrieved from http://www.whitehouse.gov/the-press-office/fact-sheet-race-top

Wyner, J. S., Bridgeland, J. M., & DiIulio Jr., J. J. (2007). Achievement trap: How America is failing millions of high-achieving students from low-income families. Jack Kent Cooke Foundation, Civic Enterprises, LLC. Retrieved from www.jkcf.org/assets/files/0000/0084/Achievement_Trap.pdf

Yonezawa, S., Wells, A. S., & Serna, I. (2002). Choosing tracks:“Freedom of choice” in detracking schools. American Educational Research Journal , 39(1), 37–67.

New report: Intelligence Explosion Microeconomics

Summary: Intelligence Explosion Microeconomics (pdf) is 40,000 words taking some initial steps toward tackling the key quantitative issue in the intelligence explosion, "reinvestable returns on cognitive investments": what kind of returns can you get from an investment in cognition, can you reinvest it to make yourself even smarter, and does this process die out or blow up? This can be thought of as the compact and hopefully more coherent successor to the AI Foom Debate of a few years back.

(Sample idea you haven't heard before: The increase in hominid brain size over evolutionary time should be interpreted as evidence about increasing marginal fitness returns on brain size, presumably due to improved brain wiring algorithms; not as direct evidence about an intelligence scaling factor from brain size.)

I hope that the open problems posed therein inspire further work by economists or economically literate modelers, interested specifically in the intelligence explosion *qua* cognitive intelligence rather than non-cognitive 'technological acceleration'. MIRI has an intended-to-be-small-and-technical mailing list for such discussion. In case it's not clear from context, I (Yudkowsky) am the author of the paper.

Abstract:

> I. J. Good's thesis of the 'intelligence explosion' is that a sufficiently advanced machine intelligence could build a smarter version of itself, which could in turn build an even smarter version of itself, and that this process could continue enough to vastly exceed human intelligence. As Sandberg (2010) correctly notes, there are several attempts to lay down return-on-investment formulas intended to represent sharp speedups in economic or technological growth, but very little attempt has been made to deal formally with I. J. Good's intelligence explosion thesis as such.
>
> I identify the key issue as *returns on cognitive reinvestment* - the ability to invest more computing power, faster computers, or improved cognitive algorithms to yield cognitive labor which produces larger brains, faster brains, or better mind designs. There are many phenomena in the world which have been argued as evidentially relevant to this question, from the observed course of hominid evolution, to Moore's Law, to the competence over time of machine chess-playing systems, and many more. I go into some depth on the sort of debates which then arise on how to interpret such evidence. I propose that the next step forward in analyzing positions on the intelligence explosion would be to formalize return-on-investment curves, so that each stance can say formally which possible microfoundations they hold to be falsified by historical observations already made. More generally, I pose multiple open questions of 'returns on cognitive reinvestment' or 'intelligence explosion microeconomics'. Although such questions have received little attention thus far, they seem highly relevant to policy choices affecting the outcomes for Earth-originating intelligent life.

The **dedicated mailing list** will be small and restricted to technical discussants.

This topic was originally intended to be a sequence in *Open Problems in Friendly AI,* but further work produced something compacted beyond where it could be easily broken up into subposts.

Outline of contents:

1: Introduces the basic questions and the key quantitative issue of sustained reinvestable returns on cognitive investments.

2: Discusses the basic language for talking about the intelligence explosion, and argues that we should pursue this project by looking for underlying microfoundations, not by pursuing analogies to allegedly similar historical events.

3: Goes into detail on what I see as the main arguments for a fast intelligence explosion, constituting the bulk of the paper with the following subsections:

- **3.1**: What the fossil record actually tells us about returns on brain size, given that most of the difference between Homo sapiens and Australopithecus was probably improved software.
- **3.2**: How to divide credit for the human-chimpanzee performance gap between "humans are individually smarter than chimpanzees" and "the hominid transition involved a one-time qualitative gain from being able to accumulate knowledge".
- **3.3**: How returns on speed (serial causal depth) contrast with returns from parallelism; how faster thought seems to contrast with more thought. Whether sensing and manipulating technologies are likely to present a bottleneck for faster thinkers, or how large of a bottleneck.
- **3.4**: How human populations seem to scale in problem-solving power; some reasons to believe that we scale inefficiently enough for it to be puzzling. Garry Kasparov's chess match vs. The World, which Kasparov won.
- **3.5**: Some inefficiencies that might cumulate in an estimate of humanity's net computational efficiency on a cognitive problem.
- **3.6**: What the anthropological record actually tells us about cognitive returns on cumulative selection pressure, given that selection pressures were probably increasing over the course of hominid history. How the observed history would be expected to look different, if there were in fact diminishing returns on cognition.
- **3.7**: How to relate the curves for evolutionary difficulty, human-engineering difficulty, and AI-engineering difficulty, considering that they are almost certainly different.
- **3.8**: Correcting for anthropic bias in trying to estimate the intrinsic 'difficulty 'of hominid-level intelligence just from observing that intelligence evolved here on Earth.
- **3.9**: The question of whether to expect a 'local' (one-project) FOOM or 'global' (whole economy) FOOM and how returns on cognitive reinvestment interact with that.
- **3.10**: The great open uncertainty about the minimal conditions for starting a FOOM; why I. J. Good's postulate of starting from 'ultraintelligence' is probably much too strong (sufficient, but very far above what is necessary).
- **3.11**: The enhanced probability of unknown unknowns in the scenario, since a smarter-than-human intelligence will selectively seek out and exploit flaws or gaps in our current knowledge.

4: A tentative methodology for formalizing theories of the intelligence explosion - a project of formalizing possible microfoundations and explicitly stating their alleged relation to historical experience, such that some possibilities can allegedly be falsified.

5: Which open sub-questions seem both high-value and possibly answerable.

6: Formally poses the Open Problem and mentions what it would take for MIRI itself to directly fund further work in this field.

Three more ways identity can be a curse

The Buddhists believe that one of the three keys to attaining true happiness is dissolving the illusion of the self. (The other two are dissolving the illusion of permanence, and ceasing the desire that leads to suffering.) I'm not really sure exactly what it means to say "the self is an illusion", and I'm not exactly sure how that will lead to enlightenment, but I do think one can easily take the first step on this long journey to happiness by beginning to dissolve the sense of one's *identity.*

Previously, in "Keep Your Identity Small", Paul Graham showed how a strong sense of identity can lead to epistemic irrationally, when someone refuses to accept evidence against x because "someone who believes x" is part of his or her identity. And in Kaj Sotala's "The Curse of Identity", he illustrated a human tendency to reinterpret a goal of "do x" as "give the impression of being someone who does x". These are both fantastic posts, and you should read them if you haven't already.

Here are three more ways in which identity can be a curse.

1. Don't be afraid to change

James March, professor of political science at Stanford University, says that when people make choices, they tend to use one of two basic models of decision making: the consequences model, or the identity model. In the consequences model, we weigh the costs and benefits of our options and make the choice that maximizes our satisfaction. In the identity model, we ask ourselves "What would a person like me do in this situation?"[1]

The author of the book I read this in didn't seem to take the obvious next step and acknowledge that the consequences model is clearly The Correct Way to Make Decisions and basically by definition, if you're using the identity model and it's giving you a different result then the consequences model would, you're being led astray. A heuristic I like to use is to limit my identity to the "observer" part of my brain, and make my only goal maximizing the amount of happiness and pleasure the observer experiences, and minimizing the amount of misfortune and pain. It sounds obvious when you lay it out in these terms, but let me give an example.

Alice is a incoming freshman in college trying to choose her major. In Hypothetical University, there are only two majors: English, and business. Alice absolutely adores literature, and thinks business is dreadfully boring. Becoming an English major would allow her to have a career working with something she's passionate about, which is worth 2 megautilons to her, but it would also make her poor (0 mu). Becoming a business major would mean working in a field she is not passionate about (0 mu), but it would also make her rich, which is worth 1 megautilon. So English, with 2 mu, wins out over business, with 1 mu.

However, Alice is very bright, and is the type of person who can adapt herself to many situations and learn skills quickly. If Alice were to spend the first six months of college deeply immersing herself in studying business, she would probably start developing a passion for business. If she purposefully exposed herself to certain pro-business memeplexes (e.g. watched a movie glamorizing the life of Wall Street bankers), then she could speed up this process even further. After a few years of taking business classes, she would probably begin to forget what about English literature was so appealing to her, and be extremely grateful that she made the decision she did. Therefore she would gain the same 2 mu from having a job she is passionate about, along with an additional 1 mu from being rich, meaning that the 3 mu choice of business wins out over the 2 mu choice of English.

However, the possibility of self-modifying to becoming someone who finds English literature boring and business interesting is very disturbing to Alice. She sees it as a betrayal of everything that she is, even though she's actually only been interested in English literature for a few years. Perhaps she thinks of choosing business as "selling out" or "giving in". Therefore she decides to major in English, and takes the 2 mu choice instead of the superior 3 mu.

(Obviously this is a hypothetical example/oversimplification and there are a lot of reasons why it might be rational to pursue a career path that doesn't make very much money.)

It seems to me like human beings have a bizarre tendency to want to keep certain attributes and character traits stagnant, even when doing so provides no advantage, or is actively harmful. In a world where business-passionate people systematically do better than English-passionate people, it makes sense to self-modify to become business-passionate. Yet this is often distasteful.

For example, until a few weeks ago when I started solidifying this thinking pattern, I had an extremely adverse reaction to the idea of ceasing to be a hip-hop fan and becoming a fan of more "sophisticated" musical genres like jazz and classical, eventually coming to look down on the music I currently listen to as primitive or silly. This doesn't really make sense - I'm sure if I were to become a jazz and classical fan I would enjoy those genres at least as much as I currently enjoy hip hop. And yet I had a very strong preference to remain the same, even in the trivial realm of music taste.

Probably the most extreme example is the common tendency for depressed people to not actually want to get better, because depression has become such a core part of their identity that the idea of becoming a healthy, happy person is disturbing to them. (I used to struggle with this myself, in fact.) Being depressed is probably the most obviously harmful characteristic that someone can have, and yet many people resist self-modification.

Of course, the obvious objection is there's no way to rationally object to people's preferences - if someone truly prioritizes keeping their identity stagnant over not being depressed then there's no way to tell them they're wrong, just like if someone prioritizes paperclips over happiness there's no way to tell them they're wrong. But if you're like me, and you *are* interested in being happy, then I recommend looking out for this cognitive bias.

The other objection is that this philosophy leads to extremely unsavory wireheading-esque scenarios if you take it to its logical conclusion. But holding the opposite belief - that it's always more important to keep your characteristics stagnant than to be happy - clearly leads to even more absurd conclusions. So there is probably some point on the spectrum where change is so distasteful that it's not worth a boost in happiness (e.g. a lobotomy or something similar). However, I think that in actual practical pre-Singularity life, most people set this point far, far too low.

2. The hidden meaning of "be yourself"

(This section is entirely my own speculation, so take it as you will.)

"Be yourself" is probably the most widely-repeated piece of social skills advice despite being pretty clearly useless - if it worked then no one would be socially awkward, because everyone has heard this advice.

However, there must be some sort of core grain of truth in this statement, or else it wouldn't be so widely repeated. I think that core grain is basically the point I just made, applied to social interaction. I.e, optimize always for social success and positive relationships (particularly in the moment), and not for signalling a certain identity.

The ostensible purpose of identity/signalling is to appear to be a certain type of person, so that people will like and respect you, which is in turn so that people will want to be around you and be more likely to do stuff for you. However, oftentimes this goes horribly wrong, and people become very devoted to cultivating certain identities that are actively harmful for this purpose, e.g. goth, juggalo, "cool reserved aloof loner", guy that won't shut up about politics, etc. A more subtle example is Fred, who holds the wall and refuses to dance at a nightclub because he is a serious, dignified sort of guy, and doesn't want to look silly. However, the reason why "looking silly" is generally a bad thing is because it makes people lose respect for you, and therefore make them less likely to associate with you. In the situation Fred is in, holding the wall and looking serious will cause no one to associate with him, but if he dances and mingles with strangers and looks silly, people will be likely to associate with him. So unless he's afraid of looking silly in the eyes of God, this seems to be irrational.

Probably more common is the tendency to go to great care to cultivate identities that are neither harmful nor beneficial. E.g. "deep philosophical thinker", "Grateful Dead fan", "tough guy", "nature lover", "rationalist", etc. Boring Bob is a guy who wears a blue polo shirt and khakis every day, works as hard as expected but no harder in his job as an accountant, holds no political views, and when he goes home he relaxes by watching whatever's on TV and reading the paper. Boring Bob would probably improve his chances of social success by cultivating a more interesting identity, perhaps by changing his wardrobe, hobbies, and viewpoints, and then liberally signalling this new identity. However, most of us are not Boring Bob, and a much better social success strategy for most of us is probably to smile more, improve our posture and body language, be more open and accepting of other people, learn how to make better small talk, etc. But most people fail to realize this and instead play elaborate signalling games in order to improve their status, sometimes even at the expense of lots of time and money.

Some ways by which people can fail to "be themselves" in individual social interactions: liberally sprinkle references to certain attributes that they want to emphasize, say nonsensical and surreal things in order to seem quirky, be afraid to give obvious responses to questions in order to seem more interesting, insert forced "cool" actions into their mannerisms, act underwhelmed by what the other person is saying in order to seem jaded and superior, etc. Whereas someone who is "being herself" is more interested in creating rapport with the other person than giving off a certain impression of herself.

Additionally, optimizing for a particular identity might not only be counterproductive - it might actually be a quick way to get people to *despise* you.

I used to not understand why certain "types" of people, such as "hipsters"[2] or Ed Hardy and Affliction-wearing "douchebags" are so universally loathed (especially on the internet). Yes, these people are adopting certain styles in order to be cool and interesting, but isn't everyone doing the same? No one looks through their wardrobe and says "hmm, I'll wear this sweater because it makes me uncool, and it'll make people not like me". Perhaps hipsters and Ed Hardy Guys fail in their mission to be cool, but should we really hate them for this? If being a hipster was cool two years ago, and being someone who wears normal clothes, acts normal, and doesn't do anything "ironically" is cool today, then we're really just hating people for failing to keep up with the trends. And if being a hipster actually *is* cool, then, well, who can fault them for choosing to be one?

That was my old thought process. Now it is clear to me that what makes hipsters and Ed Hardy Guys hated is that they aren't "being themselves" - they are much more interested in cultivating an identity of interestingness and masculinity, respectively, than connecting with other people. The same thing goes for pretty much every other collectively hated stereotype I can think of[3] - people who loudly express political opinions, stoners who won't stop talking about smoking weed, attention seeking teenage girls on facebook, extremely flamboyantly gay guys, "weeaboos", hippies and new age types, 2005 "emo kids", overly politically correct people, tumblr SJA weirdos who identify as otherkin and whatnot, overly patriotic "rednecks", the list goes on and on.

This also clears up a confusion that occurred to me when reading How to Win Friends and Influence People. I know people who have a Dale Carnegie mindset of being optimistic and nice to everyone they meet and are adored for it, but I also know people who have the same attitude and yet are considered irritatingly saccharine and would probably do better to "keep it real" a little. So what's the difference? I think the difference is that the former group are genuinely interested in being nice to people and building rapport, while members of the second group have made an error like the one described in Kaj Sotala's post and are merely trying to give off the *impression* of being a nice and friendly person. The distinction is obviously very subtle, but it's one that humans are apparently very good at perceiving.

I'm not exactly sure what it is that causes humans to have this tendency of hating people who are clearly optimizing for identity - it's not as if they harm anyone. It probably has to do with tribal status. But what is clear is that you should definitely not be one of them.

3. The worst mistake you can possibly make in combating akrasia

The main thesis of PJ Eby's Thinking Things Done is that the primary reason why people are incapable of being productive is that they use negative motivation ("if I don't do x, some negative y will happen") as opposed to positive motivation ("if i do x, some positive y will happen"). He has the following evo-psych explanation for this: in the ancestral environment, personal failure meant that you could possibly be kicked out of your tribe, which would be fatal. A lot of depressed people make statements like "I'm worthless", or "I'm scum" or "No one could ever love me", which are illogically dramatic and overly black and white, until you realize that these statements are merely interpretations of a feeling of "I'm about to get kicked out of the tribe, and therefore die." Animals have a freezing response to imminent death, so if you are fearing failure you will go into do-nothing mode and not be able to work at all.[4]

In Succeed: How We Can Reach Our Goals, Phd psychologist Heidi Halvorson takes a different view and describes positive motivation and negative motivation as having pros and cons. However, she has her own dichotomy of Good Motivation and Bad

Motivation: "Be good" goals are performance goals, and are directed at achieving a particular outcome, like getting an A on a test, reaching a sales target, getting your attractive neighbor to go out with you, or getting into law school. They are very often tied closely to a sense of self-worth. "Get better" goals are mastery goals, and people who pick these goals judge themselves instead in terms of the progress they are making, asking questions like "Am I improving? Am I learning? Am I moving forward at a good pace?" Halvorson argues that "get better" goals are almost always drastically better than "be good" goals[5]. An example quote (from page 60) is:

> When my goal is to get an A in a class and prove that I'm smart, and I take the first exam and I *don't* get an A... well, then I really can't help but think that maybe I'm not so smart, right? Concluding "maybe I'm not smart" has several consequences and none of them are good. First, I'm going to feel terrible - probably anxious and depressed, possibly embarrassed or ashamed. My sense of self-worth and self-esteem are going to suffer. My confidence will be shaken, if not completely shattered. And if I'm not smart enough, there's really no point in continuing to try to do well, so I'll probably just give up and not bother working so hard on the remaining exams.

And finally, in Feeling Good: The New Mood Therapy, David Burns describes a destructive side effect of depression he calls "do-nothingism":

> One of the most destructive aspects of depression is the way it paralyzes your willpower. In its mildest form you may simply procrastinate about doing a few odious chores. As your lack of motivation increases, virtually any activity appears so difficult that you become overwhelmed by the urge to do nothing. Because you accomplish very little, you feel worse and worse. Not only do you cut yourself off from your normal sources of stimulation and pleasure, but your lack of productivity aggravates your self-hatred, resulting in further isolation and incapacitation.

Synthesizing these three pieces of information leads me to believe that the *worst thing you can possibly do for your akrasia* is to tie your success and productivity to your sense of identity/self-worth, especially if you're using negative motivation to do so, and *especially* if you suffer or have recently suffered from depression or low-self esteem. The thought of having a negative self-image is scary and unpleasant, perhaps for the evo-psych reasons PJ Eby outlines. If you tie your productivity to your fear of a negative self-image, working will become scary and unpleasant as well, and you won't want to do it.

I feel like this might be the single number one reason why people are akratic. It might be a little premature to say that, and I might be biased by how large of a factor this mistake was in my own akrasia. But unfortunately, this trap seems like a very easy one to fall into. If you're someone who is lazy and isn't accomplishing much in life, perhaps depressed, then it makes intuitive sense to motivate yourself by saying "Come on, self! Do you want to be a useless failure in life? No? Well get going then!" But doing so will accomplish the exact opposite and make you feel miserable.

So there you have it. In addition to making you a bad rationalist and causing you to lose sight of your goals, a strong sense of identity will cause you to make poor decisions that lead to unhappiness, be unpopular, and be unsuccessful. I think the Buddhists were onto something with this one, personally, and I try to limit my sense of identity as much as possible. A trick you can use in addition to the "be the observer" trick I mentioned, is to whenever you find yourself thinking in identity terms, swap out that identity for the identity of "person who takes over the world by transcending the need for a sense of identity".

This is my first LessWrong discussion post, so constructive criticism is greatly appreciated. Was this informative? Or was what I said obvious, and I'm retreading old ground? Was this well written? Should this have been posted to Main? Should this not have been posted at all? Thank you.

---

1. Paraphrased from page 153 of Switch: How to Change When Change is Hard

2. Actually, while it works for this example, I think the stereotypical "hipster" is a bizarre caricature that doesn't match anyone who actually exists in real life, and the degree to which people will rabidly espouse hatred for this stereotypical figure (or used to two or three years ago) is one of the most bizarre tendencies people have.

3. Other than groups that arguably hurt people (religious fundamentalists, PUAs), the only exception I can think of is frat boy/jock types. They talk about drinking and partying a lot, sure, but not really any more than people who drink and party a lot would be expected to. Possibilities for their hated status include that they do in fact engage in obnoxious signalling and I'm not aware of it, jealousy, or stigmatization as hazers and date rapists. Also, a lot of people hate stereotypical "ghetto" black people who sag their jeans and notoriously type in a broken, difficult-to-read form of English. This could either be a weak example of the trend (I'm not really sure what it is they would be signalling, maybe dangerous-ness?), or just a manifestation of racism.

4. I'm not sure if this is valid science that he pulled from some other source, or if he just made this up.

5. The exception is that "be good" goals can lead to a very high level of performance when the task is easy.

Ritual Report: Schelling Day

On Sunday, April 14th, the Boston group held our first Schelling Day celebration. The idea was to open up and share our private selves. It was a rousing success.

That doesn't do it justice. Let me try again.

By all the stars, you guys. This was *beautiful*.

About fifteen people showed up. Most of us were from the hard core of Boston's rationalist community. Two of us were new to the group. (I'm hopeful this will convince them to start attending our regular meetups.) There was a brief explanation and a few vital clarifying questions before we began the ritual, which went for maybe 90-120 minutes, including a couple of short breaks. All of us spoke at least once.

I don't want to go into specifics about what people said, but it was powerful. I learned about sides of my friends I would never have guessed at. People went into depth about issues I had only seen from the surface. I heard things that will make me change my behavior towards my friends. I saw angst and guilt and hope and pain and wild joy. I saw compassion and uncertainty and courage. People said things they had never said before, things I might not have been brave enough even to *think* in their position. I had tears in my eyes more than once.

Speaking went remarkably smoothly. I set a timer for five minutes for each speaker, but it never ran out. (Five minutes is a surprisingly long time.) Partway through, Julia suggested we leave a long moment of silence between speakers, which was a very good idea and I wish I'd done a better job of enforcing it.

Afterwards, we had a potluck and mingled in small groups. At first we talked about our revelations, but over time our conversation started drifting towards our usual topics. Next time, in order to keep us on topic, I'll probably try adding more structure to this stage.

The other area I wanted to improve was the ritual with the snacks. We had five categories: Struggles, Confessions, Hopes, Joys, and Other. There weren't many Hopes, and there wasn't much distinction between Struggles and Confessions. I'll change this for next time, possibly to Hardships, Joys, Histories, and Other. There's room for improvement in the specific snacks I picked, too.

This celebration was the most powerful thing I've experienced since the Solstice megameetup. I don't think I want to do this again soon—it was one of the most exhausting things I've ever done, even if I didn't notice until after I'd left—but I know I want to do it again *sometime*.

To everyone who came: I'm so proud of what you did and who you are. Thank you for your courage and sincerity.

Four Tips for Public Speaking

TL;DR, I offered and promised in the Post Request Thread a guide to the four highest value tips I know for doing public speaking. Here they are, with explanations below:

1. ***Fortissimo!** Don't apologize for talking*
2. *Know the first and last line of your comment before you open your mouth*
3. *Think about speeches/comments as having a narrative arc*
4. *Look for additional emotional tones to layer on the content*

My background: I was a debater in college, but not in the Gatling-gun style of competitive debate. We did philosophical debate, where you only argued for propositions you *actually* believed. So, style was supposed to make it easier to get interested, but not be *too* Dark Arts-persuasive. I coached younger members on how to present their speeches and have spent a fair amount of time murderboarding people (helping people prepare for interviews or presentation).

I think the tools in this post are useful both for speeches you prepare and polish ahead of time, but also to be better at speaking coherently off the cuff (long and short form). You can check out my speaking style here. (I'm not using notes, and I didn't memorize a speech -- I memorized an *arc* which gave me room to improvise). So, here are the habits that help:

1) *Fortissimo!* Don't apologize for talking.

In E.L. Konigsberg's *About the B'nai Bagels*, the protagonist is preparing for his bar mitzvah and asks his brother for advice on how to sing his Torah portion. After listening to him, his brother has the following feedback:

> "I have only one word of advice to give you"
> "Give already"
> "That word is fortissimo… it's Italian for *loud*. When in doubt, shout, that's what I'm telling you."
> "I should shout? Everyone will hear for sure how bad I am."
> "But, my dear brother, if you sing loud and clear, it will be easier on the audience. You're making it doubly hard on them. Hard to listen to and hard to hear."

Not everyone needs to be *louder* when they speak, but a lot of people who are uncomfortable with public speaking signal that discomfort in posture or vocal tone (a lot of freshman and sophomores had an about-to-cry sounding tension in their voices when they were speaking). If you're apologizing for talking, your audience will assume there's a *reason* and start to resent it or feel uncomfortable.

So, don't apologize for talking. Don't start with disclaimers ("I'll be fast, I don't want to waste anyone's time"). And don't apologize with your voice or your body language. You can get specific feedback by taping yourself talking and have a friend watch it with you and have you practice standing taller or speaking a bit more intently. You can pay a theatre grad student to meet with you a couple times about posture of voice projection. You can also just consciously review *why* you are talking in the first place before you open your mouth, so you remember why your comment is useful and you're giving people a *gift* by talking, not being an *imposition*.

2) Know the first and last line of your comment before you open your mouth

It's pretty obvious why you want to know the first line of your speech/answer/whatever before you start talking; you don't want an awkward lag or a spot where you might panic. But most people don't plan their conclusion ahead of time (totally neglecting the peak-end rule!).

I hear a lot of novice speakers start strong, and then kind of peter out at the end of their response. Sometimes people will just trail off, hoping someone else will pick up the slack. Sometimes people have essentially already given their closing thought, but not noticed, and then they end up repeating it awkwardly.

If you know what your closing image/sentence/line/etc is when you *start* talking, you know what you're aiming at from the beginning, so you won't get diverted as easily. You've removed one common cause of failure/panic in speaking, so you can speak more confidently in the first place. And your point will be more memorable/easier to engage with if you have a strong conclusion.

3) Think about speeches/comments as having a narrative arc

So that's the opening and the closing of the talk, but what goes in the middle? In English class, you probably leaned this model:

- Thesis
- Evidence 1
- Evidence 2
- Evidence 3
- Thesis restated

This is terribly boring and difficult for people to retain. It's a lot more fun and memorable if you can put things in the framework of a story. Here's one formula for creating a narrative structure from Miss Snark:

> X is the main guy; he wants to do:
> Y is the bad guy; he wants to do:
> they meet at Z and all L breaks loose.
> If they don’t resolve Q, then R starts and if they do it’s L squared.

When I'm teaching class, I tend to use one that's more like:

> Ever notice how you always X when you'd really like to Y? So did I! I tried Z and it turned out to work, but I wasn't sure why! I poked around in the literature and found A,B, and C, which caused me to tweak my solution to Z' and now I Y all the time, and you can too!

Basically, instead of just having a point and supporting data, you take your audience through a couple emotional arcs. It's easier to remember stories than just data. It's also more fun for your listeners to repeat, so they'll get to share your idea with others. It helps you stay away from a monotone or totally even affect while speaking (more in the next tip) and keeps the structure of your comment really clear in your own head.

Planning plot summaries of my speeches means I don't need to carry notes or memorize lines, anymore than I *recite* funny stories I share with friends. I can just remember the outline of the story and then expand or contract individual parts depending on what the audience responds to. The structure gives me a safety net. This way, I'm not *unsure* what I'm saying when I open my mouth, but I'm not stuck saying specific lines.

4) Look for additional emotional tones to layer on the content

It's boring to just listen to someone explain facts. Having a narrative arc (as above) will automatically inject some variance into your tone and affect. In my teaching example above, the emotional notes look something like this:

> ***Frustration:*** [Ever notice how you always X when you'd really like to Y?] ***Shared identity, all of us looking at the frustration together:*** [So did I!] I tried Z and it turned out to work, but ***pleased but perplexed:*** [I wasn't sure why!] I poked around in the literature and ***surprise, but increasing feeling of catharsis:*** [found A,B, and C, which caused me to tweak my solution to Z'] and ***triumph:*** [now I Y all the time], ***return of fellow feeling and pleasure at sharing something cool:*** [and you can too!]

But there's more you can add. One friend of mine was explaining a counterintuitive study in a fairly matter of fact way, but it was a lot more enjoyable and memorable to hear about if she shared her surprise at how it turned out. A lot of the time, it's simplest to just make sure you're letting your honest reactions to what you're saying come across.

But, if you're not sure what those are, or want to explore other options, you can try dividing what you're saying into beats. (Beats is a phrase used in theatre for subdivisions within scenes. In one conversation or story, the dominant emotional tone can change, and that transition is the start of a new beat). So, try dividing up your notes or your outline into sections and just experiment with the dominant tone for the section. Here's a reworking of the emotional beats in my teaching outline:

> ***Sadness, regret:*** [Ever notice how you always X when you'd really like to Y?] ***Shame shared as vulnerability:*** [So did I!] I tried Z and it turned out to work, but ***tentative, a little uncertain:*** [I wasn't sure why!] I poked around in the literature and ***feeling of tinkering and assembly:*** [found A,B, and C, which caused me to tweak my solution to Z'] and ***peace, tranquility:*** [now I Y all the time], ***warmth, joy:*** [and you can too!]

Try looking at this list of some possible emotional tones, and see what it's like when you using them as you talk through your outline. Try reading *wrong* tones to a friend, to notice why they're wrong or to catch yourself if you were unnecessarily restricting your options. Sometimes tone can change a number of times in one passage (as in this marked up example), just pay attention to what prompts the shift. You can try picking a speech or a sentence that already exists, and reading it deliberately with *different* tones each time to get some practise and comfort using them.

So, if you work on these tips, people will be more comfortable listening to what you say (1), you'll open and close strongly (2), with a narrative arc that keeps you on track and makes your points memorable (3), and enough emotional variation to keep your audience engaged with you and your content (4). Huzzah!

NES-game playing AI [video link and AI-boxing-related comment]

"Pretty simple" algorithm playing games quite impressively.

http://www.youtube.com/watch?v=xOCurBYI_gY

First, this is awesome - enjoy!

Paper here http://www.cs.cmu.edu/~tom7/mario/mario.pdf

One interesting observation made by Tom Murphy is that the AI found and exploited playable bugs in the game not (commonly) known to human players. I think it's a good example to have available suggesting what a really smart AI might look for to win.