Best of LessWrong: May 2013

1. Maximizing Your Donations via a Job
2. Post ridiculous munchkin ideas!
3. The Centre for Applied Rationality: a year later from a (somewhat) outside perspective
4. The Power of Pomodoros
5. Pascal's Muggle: Infinitesimal Priors and Strong Evidence
6. Robustness of Cost-Effectiveness Estimates and Philanthropy
7. 10-Step Anti-Procrastination Checklist
8. LW Study Hall - 2 Month Update
9. Wikifying the blog list
10. Be Nice to Non-Rationalists
11. What do professional philosophers believe, and why?

Best of LessWrong: May 2013

1. Maximizing Your Donations via a Job
2. Post ridiculous munchkin ideas!
3. The Centre for Applied Rationality: a year later from a (somewhat) outside perspective
4. The Power of Pomodoros
5. Pascal's Muggle: Infinitesimal Priors and Strong Evidence
6. Robustness of Cost-Effectiveness Estimates and Philanthropy
7. 10-Step Anti-Procrastination Checklist
8. LW Study Hall - 2 Month Update
9. Wikifying the blog list
10. Be Nice to Non-Rationalists
11. What do professional philosophers believe, and why?

Maximizing Your Donations via a Job

In November of 2012 I set a goal for myself: find the most x-risk reducing role I can fill. At first I thought it would be by working directly with MIRI, but after a while it became clear that I could contribute more by simply donating. So my goal became: find the highest paying job, so I can donate lots of money to CFAR and MIRI.

A little bit of background on me. Started programming in 2000. Graduated in 2009 with Bachelor's in computer science. Worked for about a year and a half at a game company. Then did my own game startup for about a year. Then moved to the bay area and joined a game startup here, which was acquired 10 months later. Worked a bit at the new company and then left. So, just under four years of professional programming experience, but primarily in the game industry. Almost no leadership / managerial experience, aside from the startup I did where I hired freelancers.

Below is my experience of finding a software engineering job in the Silicon Valley. If you are not an engineer or not in the Silicon Valley, I think you'll still find a lot of useful information here.

Pre-game

Before sending out my resume, I spent about a month preparing. I read Intro to Algorithms, which was very good overall, but not a huge help in preparing for interviews.[1] I read Cracking the Coding Interview, which was extremely helpful. (If you read only one book to prepare, make it this one.) The book has a lot of questions that are similar to the ones you'll actually see during interviews. I also did TopCoder problems, which were pretty helpful as well.[2] Looking back, I wish I spent more time finding actual interview questions online and doing more of those (that's why CCI book was so helpful).

After several weeks of preparation, I compiled a long list of companies I was going to apply to. I checked on GlassDoor to see what kind of salary I could expect at each one. I then rated all the companies. Companies with low salaries and poor personal fit received the lowest rating.

I started by applying to companies with the lowest ratings. This way I could use them as practice for the companies I thought would actually make a competitive offer. This was the right move and worked very well. (Another friend of mine did the same approach with good results as well.) Remember, you are not just doing those interviews to practice the coding problems, you are practicing pitching yourself as well.

Interviewing with a company

Standard procedure for applying to a tech company:

1. Send them your resume.

- Proofread your resume. Let your friends proofread it.
- Make sure there are only relevant things on it. When I applied to tech companies, I removed a lot of game-specific things from my resume. When I applied to companies that did 3D graphics, I made sure I had all my 3D graphics experience listed. I ended up with two version of my resume.
- Have your resume in DOC, PDF, and TXT formats. This way you'll always have the right one when you upload / paste it.
- For a few companies, I had a friend or friend of a friend who referred me. This REALLY HELPS in two ways: 1) your resume will be processed a lot faster, 2) if your friend is a great engineer/employee, you'll be taken a lot more seriously, and the company will fight for you a lot harder.

2. You'll get an email from the recruiter and setup a time to speak, where you'll talk about yourself, what you've done, why you are interested in their company, and so on. You can and should ask them questions as well.

- When you start getting multiple calls each day, make sure you know who is calling. There is nothing worse than talking about the challenges of streaming music to a car sharing startup. (True story.)
- Read about the company on Wikipedia before the call. Know the basic stuff. Look at their website and read the About page.
- Find the thing that makes the company special and successful. Find the thing that you actually think is cool about the company. Those are your answers for why you want to work there.
- Ask non-technical questions: How is the company structured? How many teams are there? How many employees? Engineers? Think of other intelligent questions to ask.
- In my experience, it's not very beneficial to tell them you are interviewing with a dozen other companies. When they ask who else you are interviewing with, just name a few companies, especially the competitors / similar companies.
- Be SUPER NICE to your recruiter. They are your main point of contact with the company. They'll be the one fighting to get you the best offer.

3. You'll have a technical phone interview with a software engineer where you'll solve a problem or two on collabedit or some similar website. At the end, you'll get a few minutes to ask them questions too.

- All the usual interviewing tips apply here. E.g. talk out loud, your interviewer doesn't know what you are thinking.
- Most companies don't care what language you use, as long as it's mainstream. (I used C# for almost all my coding questions.)
- DO NOT start answering the question by writing code. If the questions seems vague, ask about the context. Who'll be using this solution? Definitely ask about the kind of data you are working with. If it's integers, are they random? Over some small range or over all possible integers?
- List out metrics for various approaches: brute-force solution, optimized for speed solution, optimized for memory solution. Here is a question I saw a few times: Write a data structure which can accept and store integers, and can check if

there exist two integers that sum up to a given number. There are multiple solutions, and the best one depends on the ratio of addInteger to checkForSum calls.
- The previous steps should only take you a minute or two. Once you've decided what the best approach is, then you can write the solution. When you are done, check for errors, then run through several examples. Do a simple example and a slightly complicated example. When you find a bug, don't be hasty in fixing it. Understand why it happened and make sure you won't introduce new bugs by fixing it.
- If everything works, make sure you handle errors correctly. Can you handle invalid input? Input that violates your assumptions? (As a reminder, I leave “\\Check for errors” comments in appropriate spots as I code the solution.)
- When you are done, ask the interviewer questions. Ask them to tell you about what they do, if they haven't already. What have they been working on recently? What technologies/languages do they use at the company? Do they use Scrum/Agile? Pair-programming? Come up with other intelligent questions to ask.

4. You'll be invited for an on-site interview which will be 3-6 hours long, at least half of which will be coding on a white-board. (Although, a friend told me he brought his laptop with him, and most people were fine with him coding on it.)

- All the previous tips apply.
- Be on time. Take bathroom breaks when you need them. I found that drinking water during the interview keeps me refreshed. Remember your posture, body-language, and eye-contact skills.
- Learn how to talk out loud as you are writing out your solution. If you are stuck, explain what you are thinking, and what your intuition is telling you.
- Learn how to read your interviewers. If you say, "Here we should check for the null case or for empty array," and they go "Yeah, yeah, okay," they are not the type of interviewer that really cares about error conditions, so you can be somewhat more lax there. By the time I was finishing my on-site interviews, I could tell if my solution was right just by the interviewer's body language.
- When you are done, ask them questions. What are they working on? What's the thing they like most about the company? What's their least favorite thing about the company? (Another way to phrase that: What's one thing you would change if you could change anything about the company?) Do they have to work overtime? How are the people here? Can you switch between projects? Are there company wide events? In all my interviews I've never met an interviewer that didn't try to sell their company really hard. People will always tell you their company is the best place to work.
- If the person is a manager or a director, ask them higher level questions. What kind of culture are they trying to create? What are the current big challenges? Where do they want the company to be in the next 5 years? How does one advance in the company? (Usually there is a managerial and a technical track.) How often are reviews done? How are they structured?

5. You'll get a call from the recruiter congratulating you on an offer. They'll go over the offer details with you.

- Before they make you an offer, they'll check if you are actually seriously considering their company. If you told a startup you are also interviewing with Google, they might suspect that you are not seriously considering them. Unless you dissuade those fears, they might actually not even make you an offer. (Happened to me with Rdio.)
- If you didn't get an offer, try to get as much info as you can. What happened? What can you improve on? Below are the reasons why I didn't get an offer after an on-site interview:
    - Not doing well on a technical question. (Happened twice; one time because of a very obnoxious interviewer.)
    - Not interviewing for quite the right position (that on-site interview ended early).
    - Not having the necessary experience (a lot more important to startups than bigger companies).
    - Not being passionate enough about the company.
- If this is not a good timing for the offer, e.g. it's one of your first interviews, then tell them so. They will probably wait to give you the offer details until you are ready to consider it.
- The recruiter will likely ask what's important to you in an offer. How are you going to make your decision? What I've said is that compensation will be an important factor in my decision, but that the team/project/etc. are important considerations as well.

6. You have a few days (usually around 5 business days) until the offer expires to decide if you want to accept it.

- Sometimes the offer will expire before you've received offers from other companies. This is why it's important to interview in rough order of ranking, so that you can just let those offers go, knowing you'll have much better ones soon. If you want to hold on to the offer, just ask your recruiter for an extension. It'll be much easier to get an extension at big companies, especially if you are interviewing for a generic position.
- If you decline the offer, let them know.

Always be very nice, friendly, and polite. Walk the fine line between telling the truth and saying the right thing. Ideally, make sure those are the same. Even if you are interviewing with a company you have no intention of working at, make sure to find something you really like about them, something that makes them stand out to you. Always have a good answer to: "Why do you want to work here?"

Before each on-site interview make sure you research the company thoroughly. Use their product. Think of ways to improve it. It's very helpful if you can meet with someone that works there and talk to them. See if they can give you any tips on the interview process. Some companies (e.g. AirBnB) want people that are *extremely* passionate about their product. Some companies focus more than usual on architectural questions. Many companies expect the engineers to have some familiarity with UI/UX and the ability to think about a feature from all angles.

Managing your time

I sent my resume to 78 companies, had at least a phone conversation with a recruiter with 27 of them, had an on-site interview with 16 companies, and received 12 offers. Out of those, I've only seriously considered 3. (Companies with lower ratings had an atrocious response rate.)

My time-line ended up looking something like this:

- Week 1: Started applying to low-rated companies. About 2 phone interviews.
- Week 2: About 7 phone interviews. One on-site interview. Sending out more resumes.
- Week 3: About 3 phone interviews.
- Week 4: About 15 phone interviews. A few meetings with friends of friends, who ended up referring me. 1 on-site interview. Sent my resume to all the high-rated companies. (During this week interviewing became a full-time job.)
- Week 5: About 10 phone interviews. 4 on-site interviews.
- Week 6: 8 phone interviews. 4 on-site interviews.
- Week 7: 4 phone calls. 5 on-site interviews.
- Week 8: 12 phone calls. 2 on-site interviews.
- Week 9: About 8 calls a day for a few days, while I negotiated with my top companies.
- (These are strictly lower bounds for phone calls. On-site data is pretty accurate.)

Some companies move fast, some companies move slow. Google took 2 weeks from the on-site interview to the offer call. This is very common for them, but most other companies move faster. With Amazon, I actually interviewed with two different branches. With one branch things were going well, until they dropped the ball and never got back to me, even after I pestered them. This is unusual; although, Twitter did something similar, but then ended up responding with an on-site invitation. With the other Amazon branch, when I got home from the on-site interview, I already had an email saying they were going to make an offer. This is extremely fast. (I had a very good reference for that position.) Most companies take about a week between on-site and offer. The whole process, from first call to offer, takes about three weeks.

If your recruiter doesn't respond to you during 4 days or longer, shoot them an email. They might have forgotten to respond, or thought they did, or may be things are moving slowly, or may be they decided not to pursue. You want to be clear on where you stand with all the companies you are applying to.

The timing is pretty important here. You want your top-rated companies to give you an offer within a span of a week. This way you'll be able to leverage all those offers against each other.

If your current job position is already almost optimal for your goals, then it's possible you can do a few interviews, get a few offers and pick the best one, which will give you some marginal improvement. Or use those offers to leverage a raise at your existing company. But if you are pretty sure your current job has not been optimized for your goal, then I'd say, contrary to popular wisdom, just leave and spend a full month interviewing. (Or, even better, if you can, take a long "vacation".) You just can't do this kind of intense interviewing while holding another job. The one exception to this rule I can think of is if one of your highest-rated companies is a competitor with your current employer. Then you can leverage that!

Value of information is extremely high during this process. Talk to all the companies you can, talk to all the people you can. Once you have the final list of companies you are considering, reduce your uncertainty on everything. Validate all your assumptions. (Example: I was *sure* Google matched donations up to $12k, but turns out it's only up to $6k.)

How to evaluate your offer

There are 4 basic components in an offer: sign-on bonus, base salary, equity, and bonus.

Sign-on bonus. Most companies will be okay offering something like $12k sign-on bonus. Some will offer more. Most startups probably won't offer any.

Base salary. This is pretty consistent across most companies. Based on your experience, you'll be given a title (e.g. Senior Software Engineer or SE 2), and that title will determine the range of the salary you can expect. If you are good, you can demand a salary at the top of that range, but it's extremely hard to go higher.

Equity. This is the most interesting part. A good amount of value will come from this portion. With a startup, it'll be most of it. Here are two things to pay attention to:

- Is the company public or private? If it's public, you are most likely going to be given RSUs (restricted stock units), which will basically convert to normal company shares when they vest. For private companies, see the section below.
- What's the vesting schedule? For almost all companies you'll get 25% of your shares right after your first year. (This is called a 'cliff'.) After that you'll be given the appropriate fraction either monthly (e.g. at Google) or quarterly (e.g. at Facebook). Amazon is an example of a company where the vesting schedule is somewhat different: 5% after year 1, 15% after year 2, and then 20% each semester for the next two years.

Bonus. This is the bonus system the company has setup. You can't negotiate it, but it's important to take it into account.

- There will usually be a cash bonus that's based on your salary. It'll have a target percent (e.g. 15%). If you can find out how many people hit their target, that will be very helpful. However, most companies don't share or simply don't have that information.
- Some companies also have equity bonuses. Try to get as much info on those as you can. Don't assume that you'll get the maximum bonus even if you work hard. If you have friends working at that company, ask them what kind of bonuses they've been getting.
- Lots of startups don't have bonus systems in place.

Other factors.

- Donation matching: Google matches up to $6k (you donate $6k to any charity, they'll donate another $6k). Craigslist matches 3:1 up to 10% of your salary. Most companies don't have anything like that, and you can't negotiate it.
- Paid Time Off: Google offers 2 weeks, all other companies I was considering offer 3 weeks, and some even have unlimited PTO. This is not negotiable in most

companies.
- Commute: how far will you have to travel to work? Are you okay moving closer to work? (Google and Facebook have shuttles that can pick you up almost anywhere, so you could work while you commute.)
- People/culture/community/team/project are all important factors as well, depending on what you want. If you are going to spend the next several years working on something, you should be building up skills that will still be valuable in the future.

Thinking about private companies

If the company is private, you might be given RSUs or you might be given stock options. With stock options, you'll have to pay the strike price to exercise your options. So the total value your options have is: (price of a share - strike price) * number of shares.

You can't do anything with your shares until the company gets acquired or goes public. Some companies have liquidation events, but those are pretty rare. Most companies don't have them, and the ones that do only extend the opportunity to people that have been with the company for a while. There are also second-hand markets, but I don't know much about those.

If you are completely risk-intolerant, then just go with a public company, and don't consider private companies. (This is actually not exactly true. Just because a company is public, doesn't mean its risk-free, and just because a company is private doesn't mean there is a lot of risk. There are other important factors like the size of the company, their market diversity, and how long they've been around.) If you are okay with some risk, then you want a company that's close to an IPO or is likely to get acquired soon. If you want to have a chance to make more than a few million dollars, either start your own company or join a very early stage startup (my top pick would be Ripple). Before doing so, check out the stats on startups to make sure you understand how likely any given startup is to fail and make sure you understand the concepts of inside/outside view.

Taxes

It's crucial to understand all the tax implications of your salary, equity, and donations. I'm not going to go into all the details, there are a lot of resources out there for this, but you should definitely read them until it's crystal clear how you will be taxed. I'll highlight a few points:

- Understand the tax rate schedule and notice the new 39.6% tax bracket. If your income is $100k, that doesn't mean you get taxed 28% on all of it. 28% applies only to the income portion above $87,850. Also note that this is only the federal tax. Your state will have additional taxes as well. Aside from those percentages, there are a few other flat taxes, but they are considerably smaller in magnitude.
- The money you donate to a nonprofit (aka. 501(c)(3)) organization can be subtracted from your taxable income. This means that you will most likely get a

refund when you file your taxes. Why? Because when you fill out your W4 form, you'll basically tell your employer how much money to withhold from your paycheck for tax purposes. If you don't account for your future donations, more money will be withheld than is appropriate and the discrepancy will be paid back to you after you file your taxes. Ideally, you want to take your donations into account and fill out the W4 form such that there are no discrepancies. That means you'll get your money now rather than later. (I haven't gone through this process myself, so there is some uncertainty here.)

- You can claim tax deduction for up to 50% of your wages. That means if you make a lot of money in one year, even if you donate most of it, you'll be able to reduce your taxable income by a maximum of 50%. The rest goes over to the next year.
- When RSUs vest, their value is treated as ordinary income for tax purposes. When you sell them, the difference is taxed as a capital gain (or loss).
- Stock options have a more complicated set of tax rules, and you should understand them if you are considering a company that offers them.
- You can't have your employer donate money or stock for you to bypass the taxes. I've asked.

Calculating donations

To calculate exactly how much I could donate if I worked at a given company, I've created this spreadsheet. (This is an example with completely fictitious company offers with very low numbers, but the calculations should be correct.) Let me walk you through the spreadsheet.

Time discounting (Cell B1)

Money now is more valuable than money later. By how much? That's a very complicated question. If you invest your money now, you might be able to make something like 10% annually with some risk.[3] If you are donating to a charity, and they are growing very rapidly, then they can do a lot with your money right now, and you should account for that as well. If you expect the charity to double in size/effectiveness/output in the next year, then you might use a discount rate as high as 50%. I chose to use 20% annual discount rate based on my own estimates. Since I'm doing monthly compounding, the spreadsheet value is slightly higher (~22%). You can look at the column K to see how the future value of a dollar is being discounted. Note, for example, that a dollar in 12 months is worth 80¢ to me now. This discounting rate is especially important to keep in mind when examining startups, because almost all their compensation lies in the future. The further away it is, the more heavily you have to discount it.

Cost of living (Cell B2)

This is how much pre-tax money a year I'm not going to donate. See column L for the monthly expenses. We time-discount those dollars as well.

Offers (Cells A4-I15)

This is where you plug-in the offers you get. *Bonus* row is for cash bonus. *Equity* row is for the total equity the company offers you. I use the dollar amount, but you'll notice that for some of them I'm computing the dollar amount as: RSUs the company is giving me * current share price. For private companies, this is value I expect my equity to have when the company goes public. For Square it looks like: (percent of the company I'll own) * (my guess at valuation of the company at IPO) - (cost to exercise my options). For Twitter it looks like: (growth factor up to IPO) * (current price per share) * (RSUs I am granted). (Again, the numbers are completely made up.) In my calculations I'm not expecting public companies' share price to rise or fall. If you disagree, you should adjust for that as well.

Monthly projections (Cells A18-I66)

We are going to look at how much money we'll be making per month for the next four years. (Four years because our equity should be fully vested by that time.) If you are certain that you will stay at the company for less time than that, then you should consider a shorter timeline. This might affect companies differently. For example, most of the equity you get at Amazon comes during the last two years. If you are not going to be there, you are missing out on a big part of your offer.

For companies that I was seriously considering, I created two columns: one for cash wages and one for equity wages. This way I can do taxes on them more precisely.

Let's go through the Google's offer:

- For the first year we'll be only making our standard salary.
- After the first year, we get our cash bonus (green font). Here we are assuming it'll be 15% of our salary. We also get 25% of our RSUs vested (salmon background).
- For the remainder of the second year, we are making our normal salary. Each month we also get $1/48^{th}$ of our original equity offer.
- Google also has an equity bonus system, where each year you can get a bonus of up to 50% of your original equity offer. This bonus will be paid in RSUs, and it vests over 4 years, but with no cliff. So we count that as well, but I'm assuming I'm only going to get 15%, not the full 50%.
- In year 3 everything is basically the same, except now we got our second equity bonus, so we have two of them running simultaneously.
- In year 4, we have three of them running simultaneously.

For pre-IPO companies, I've estimated when they'll go IPO. Most have clauses in place that don't allow you to sell your shares until after half a year or so after the IPO. I'm assuming I will sell/donate all my shares then, and then continue selling/donating them as they continue vesting.

Sum (Cells A68-I71)

In row 68 we have the total sum. This is the amount of pre-tax dollars we expect to earn in the next four years (remember that this amount has been adjusted for time-discounting, so it'll seem much lower than you'd normally expect). L68 is how much money we are spending on ourselves during those four years.

In row 69 we subtract our living expenses to get the amount of money we'll be able to donate. Note that I'm subtracting it from the cash column, leaving the equity column alone (for the companies where I split the two).

In row 70 we account for taxes. Note that our living expenses already accounted for the taxes we pay up to $65k, so the rest of it will be taxed at around 28% or higher. You could sell your shares, or you could just donate your shares directly to your charity. (That's what we are doing with our Google offer.)

In row 71 we simply sum up the donations from cash and equity.

Disclaimer 1: while I tried as hard as I could to double check this spreadsheet, there might still be mistakes there, so use it with caution and triple check everything. The tax calculations as they are right now are wrong, and you'll have to redo them (basically the whole Row 70) based on your own numbers.

Disclaimer 2: this spreadsheet is not great for evaluating an offer from a startup, since it doesn't capture the associated uncertainty and risk. Furthermore, if you expect the startup to succeed after more than 4 years, to correctly compare it to other companies you'll have to compute more than 48 months and potentially start accounting for things like promotions and raises.

Picking the one

All right, so how do you actually pick the best company? It's not as simple as picking the one with the highest EV, since you have to account for risk involved with startups and even pre-IPO companies. In fact, you should be surprised if your offers from public companies have a higher EV than offers from startups. If that's the case, I'd double check your calculations.

This is where it becomes extremely crucial to narrow down your uncertainty. When is the company going to IPO? What is the likely valuation? Does the company have a lot of competitors? Does the company have the necessary talent to execute on their plan? What's the company's history? What is the employee churn rate (especially for executives)? How well is the company doing financially? Who are the investors? Etc, etc, etc... There is a ton of questions you should be asking, and you should be asking them to everyone whose opinion on this issue you can respect. Honest opinion from an informed and knowledgeable neutral party is worth a LOT here!

You should also talk to the people at the company. Your recruiter will connect you to the right people if you ask. Keep in mind that nobody there will tell you that the

company is going to go bankrupt or fail. But you can still get some valuable estimates, and then potentially discount them down a bit. You can even ask for their opinion on other companies you are interviewing with. Expect them to completely throw the other company under the bus though, but even so, you could get a lot of valuable criticism and bring it up when you talk to that other company. Overall, expect a lot of conflicting messages.

Keep in mind the charities you'll be donating to. What kind of donors do they have already? Are most people donating a bit from their salary? In that case, a more risky venture might be reasonable. Can they really use some money right now, or would they be a lot more effective later on with a large capital? What's their time discount rate? If you care about your charity, you can help them diversify their donor pool.

For me, it was a hard choice between big public companies (primary candidate: Google) and close to IPO companies (primary candidates: Twitter and Square).

Negotiating

You have to negotiate your offer. You have to *have to* **have to** HAVE TO. For any given company, you'll be able to get them to up their offer at least once and potentially thrice. Example: Google upped my offer three times.

- Some companies will tell you their offer is not negotiable. That's not true.
- It's much easier to leverage similar companies against each other. Leverage big public companies against each other; leverage pre-IPO companies against each other; etc... Leveraging between those categories is a bit more difficult, because startups know they can't compete with the raw cash value you are offered at bigger companies. The only thing they can do is up their equity offer and hope that they are a much better personal fit for you than the large companies.
- Recruiters will ask you very directly what the other companies are offering you. You can choose to disclose or not to disclose. If you don't disclose, the company will come back to you with their standard offer. That offer might be higher or lower than you expected. (Example: The first offer I got from Google was significantly worse than initial offers I got from Facebook and Amazon.) If you tell them what offers you have (and you should only disclose details of your very best offers), then they'll very likely match or come in a bit stronger. Usually you don't have much to gain by disclosing your other offers upfront. You can always do so later. However, you should let your recruiters know that other companies *did* make an offer, or you are expecting them to. That gives you more leveraging power.
- Sign-on bonus is very easy to negotiate. You can easily convince a company to match a sign-on bonus their competitor has offered.
- Negotiating salary is much harder, but, again, usually you can convince a company to match a salary their competitor has offered or at least come closer to it. If you are interviewing with startups, their salary offer will usually be lower than at bigger companies and even harder to negotiate. ("Cash is king" is the common phrase used there.)

First negotiating phase: simply email / call back your recruiter (who is now your best friend, right?) and tell them that the offer is somewhat lower than you expected, you have other better offers from other companies, and you are wondering if they can

increase their offer. If the company made you a clearly worse offer than another similar company, you should be very open about it.

Second negotiating phase: matching other companies. This is when it makes the most sense to disclose your other offers. For example, I used my Amazon and Facebook offers to convince Google to up their offer significantly. For some reason their original offer was very low, but seeing their competitors with much better offers convinced them to update pretty quickly. You can also bring up the perks one company has that the other doesn't (e.g. donation matching or unlimited PTO). The company can make up for that with salary/equity. There is some difficulty in using offers from private companies as leverage, because there is not much information you can disclose about them. You can talk about the number of shares you'll have, but it might not mean anything to the other recruiters if they are not familiar with the startup.

Third negotiating phase: once you picked the company you'll work for, go back to them and say something along the lines of "I really like the offer and the company, but there are a few things that don't make it ideal for me. One of your competitors did this, and another company has that. Right now I'm inclined to go with your competitor, but it's a tough decision, and I would rather go with you. I think if you can make me an offer with the following parameters, it'll make my decision extremely easy, and I'll sign on the spot." Include offer letters from other companies, especially the ones that have them beat or beat on some parameters. Notice the key promise at the end: *you will sign with them*. Your recruiter will have a lot more leverage in fighting for you if you make that promise. You are not legally obligated to follow through with your promise, but I wouldn't advise breaking it or using it just to extract more value to use as leverage against other companies. Use this tactic at the very end to extract that last bit of value from the company that's already the best. This is what I did with Google. I asked for about 3% higher salary and 12% more equity than what they were offering, and they came back with the exact numbers I requested, which means I should have asked for more. My advice would be to ask for about twice or may be even three times as much (6% and 30% respectively). Even if they come back with a compromise, it'll very likely be more than 3% and 12% increase. If not, you can try to barter one more time.

I'm sure some people will cringe at this kind of haggling, but, in all honesty, this is what recruiters expect, and they are very much used to it. Nobody even blinked an eye when I started negotiating, even on second and third rounds. However, some recruiters might try to make you feel guilty. They'll say that if you really want to work at their startup, then you shouldn't really care about your compensation. Most points they'll make will even be valid, but if you are trying to optimize for donations, then you have to make the compensation the most important factor in your decision. I've actually told most of my recruiters that I plan to donate most of my salary to charities. I don't think that got me higher offers, but it made me come off less like a greedy jerk.

At the end of the day, the company wants you, but they want to pay you as little as possible. *But*, given the choice of having you and paying you the most you deserve VS. not having you, all companies will pick the first option. ALL OF THEM. This is one of the best perks of being a talented software engineer in the bay area.

Once you accept the offer, don't forget to email everyone else and let them know. Thank everyone that helped you. Some recruiters will be surprised by your decision, and some will even fight really hard to get you to reconsider.

[1] None of the interviews required a data structure more complicated than a heap. All the answers had a very easy to compute complexity, either polynomial, polynomial * logarithmic, or factorial. The most weird one was probably O(√n) for computing prime numbers.

[2] Some problems I did during actual single-round match-up (SRM) competitions, which is good for training yourself how to code and think faster than you are used to. I also did a lot of old SRM problems, which have solutions and explanations posted in case I couldn't get them. I could easily do problem 1 & 2 in the easy division, and could do problem 3 most of the time. I didn't really bother with the hard division, and none of the interview questions were ever as hard as problem 3 in the easy division.

[3] According to the comments, this number is too high. Pick your own best estimate.

Post ridiculous munchkin ideas!

Thus spake Eliezer:

> A Munchkin is the sort of person who, faced with a role-playing game, reads through the rulebooks over and over until he finds a way to combine three innocuous-seeming magical items into a cycle of infinite *wish* spells. Or who, in real life, composes a surprisingly effective diet out of drinking a quarter-cup of extra-light olive oil at least one hour before and after tasting anything else. Or combines liquid nitrogen and antifreeze and life-insurance policies into a ridiculously cheap method of defeating the invincible specter of unavoidable Death. Or figures out how to build the real-life version of the cycle of infinite *wish* spells.

It seems that many here might have outlandish ideas for ways of improving our lives. For instance, a recent post advocated installing really bright lights as a way to boost alertness and productivity. We should not adopt such hacks into our dogma until we're pretty sure they work; however, one way of knowing whether a crazy idea works is to try implementing it, and you may have more ideas than you're planning to implement.

So: please post all such lifehack ideas! Even if you haven't tried them, even if they seem unlikely to work. Post them separately, unless some other way would be more appropriate. If you've tried some idea and it *hasn't* worked, it would be useful to post that too.

The Centre for Applied Rationality: a year later from a (somewhat) outside perspective

I recently had the privilege of being a CFAR alumni volunteering at a later workshop, which is a fascinating thing to do, and put me in a position both to evaluate how much of a difference the first workshop actually made in my life, and to see how the workshops themselves have evolved.

Exactly a year ago, I attended one of the first workshops, back when they were still inexplicably called “minicamps”. I wasn't sure what to expect, and I especially wasn't sure why I had been accepted. But I bravely bullied the nursing faculty staff until they reluctantly let me switch a day of clinical around, and later stumbled off my plane into the San Francisco airport in a haze of exhaustion. The workshop spat me out three days later, twice as exhausted, with teetering piles of ideas and very little time or energy to apply them. I left with a list of annual goals, which I had never bothered to have before, and a feeling that more was possible–this included the feeling that more would have been possible if the workshop had been longer and less chaotic, if I had slept more the week before, if I hadn't had to rush out on Sunday evening to catch a plane and miss the social.

Like I frequently do on Less Wrong the website, I left the minicamp feeling a bit like an outsider, but also a bit like I had come home. As well as my written goals, I made an unwritten pre-commitment to come back to San Francisco later, for longer, and see whether I could make the "more is possible" in my head more specific. Of my thirteen written goals on my list, I fully accomplished only four and partially accomplished five, but I did make it back to San Francisco, at the opportunity cost of four weeks of sacrificed hospital shifts.

A week or so into my stay, while I shifted around between different rationalist shared houses and attempted to max out interesting-conversations-for-day, I found out that CFAR was holding another May workshop. I offered to volunteer, proved my sincerity by spending 6 hours printing and sticking nametags, and lived on site for another 4-day weekend of delightful information overload and limited sleep.

Before the May 2012 workshop, I had a low prior that any four-day workshop could be life-changing in a major way. A four-year nursing degree, okay–I've successfully retrained my social skills and my ability to react under pressure by putting myself in particular situations over and over and over and over again. Four days? Nah. Brains don't work that way.

In my experience, it's exceedingly hard for the human brain to do *anything* deliberately. In Kahneman-speak, habits are System 1, effortless and automatic. Doing things on purpose involves System 2, effortful and a bit aversive. I could have had a *much* better experience in my final intensive care clinical if I'd though to open up my workshop notes and tried to address the causes of aversions, or use offline time to train habits, or, y'know, do *anything* on purpose instead of floundering around trying things at random until they worked.

(The again, I didn't apply concepts like System 1 and System 2 to myself a year ago. I read 'Thinking Fast and Slow' by Kahneman and 'Rationality and the Reflective Mind' by Stanovich as part of my minicamp goal 'read 12 hard nonfiction books this year', most of which came from the CFAR recommended reading list. If my preceptor had had any idea what I was saying when I explained to her that she was running particular nursing skills on System 1, because they were engrained on the level of habit, and I was running the same tasks on System 2 in working memory because they were new and confusing to me, and that was why I appeared to have poor time management, because System 2 takes forever to do anything, this terminology might have helped. Oh, for the world where everyone knows all jargon!)

...And here I am, setting aside a month of my life to think only about rationality. I can't imagine that my counterfactual self-who-didn't-attend-in-May-2012 would be here. I can't imagine that being here now will have *zero* effect on what I'm doing in a year, or ten years. Bingo. I did one thing deliberately!

So what was the May 2013 workshop actually like?

The curriculum has shifted around a lot in the past year, and I think with 95% probability that it's now more concretely useful. (Speaking of probabilities, the prediction markets during the workshop seemed to flow better and be more fun and interesting this time, although this may just show that I was more averse to games in general and betting in particular. In that case, yay for partly-cured aversions!)

The classes are grouped in an order that allows them to build on each other usefully, and they've been honed by practice into forms that successfully teach skills, instead of just putting words in the air and on flipcharts. For example, having a personal productivity system like GTD came across as a culturally prestigious thing at the last workshop, but there wasn't a lot of useful curriculum on it. Of course, I left on this trip wanting to spend my offline month creating with a GTD system better than paper to-do lists taped to walls, so I have both motivation and a low threshold for improvement.

There are also some completely new classes, including "Againstness training" by Valentine, which seem to relate to some of the 'reacting under pressure' stuff in interesting ways, and gave me vocabulary and techniques for something I've been doing inefficiently by trial and error for a good part of my life.

In general, there are more classes about emotions, both how to deal with them when they're in the way and how to use them when they're the best tool available. Given that none of us are Spock, I think this is useful.

Rejection therapy has morphed into a less terrifying and more helpful form with the awesome name of CoZE (Comfort Zone Expansion). I didn't personally find the original rejection therapy all that awful, but some people did, and that problem is largely solved.

The workshops are vastly more orderly and organized. (I like to think I contributed to this slightly with my volunteer skills of keeping the fridge stocked with water bottles and calling restaurants to confirm orders and make sure food arrived on time.) Classes began and ended on time. The venue stayed tidy. The food was excellent. It was easier to get enough sleep. Etc. The May 2012 venue had a pool, and this one didn't, which made exercise harder for addicts like me. CFAR staff are talking about solving this.

The workshops still aren't an easy environment for introverts. The negative parts of my experience in May 2012 were mostly because of this. It was easier this time, because as a volunteer I could skip classes if I started to feel socially overloaded, but periods of quiet alone time had to be effortfully carved out of the day, and at an opportunity cost of missing interesting conversations. I'm not sure if this problem is solvable without either making the workshops longer, in order to space the material out, and thus less accessible for people with jobs, or by cutting out curriculum. Either would impose a cost on the extroverts who don't want an hour at lunch to meditate or go running alone or read a sci-fi book, etc.

In general, I found the May 2012 workshop too short and intense–we had material thrown at us at a rate far exceeding the usual human idea-digestion rate. Keeping in touch via Skype chats with other participants helped. CFAR now does official followups with participants for six weeks following the workshop.

Meeting the other participants was, as usual, the best part of the weekend. The group was quite diverse, although I was still the only health care professional there. (Whyyy???? The health care system needs more rationality *so* badly!) The conversations were engaging. Many of the participants seem eager to stay in touch. The May 2012 workshop has a total of six people still on the Skype chats list, which is a 75% attrition rate. CFAR is now working on strategies to help people who want to stay in touch do it successfully.

Conclusions?

I thought the May 2012 workshop was awesome. I thought the May 2013 workshop was about an order of magnitude more awesome. I would say that now is a great time to attend a CFAR workshop...except that the organization is financially stable and likely to still be around in a year and producing even *better* workshops. So I'm not sure. Then again, rationality skills have compound interest–the value of learning some new skills now, even if they amount more to vocab words and mental labels than superpowers, compounds over the year that you spend seeing all the books you read and all the opportunities you have in that framework. I'm glad I went a year ago instead of this May. I'm even more glad I had the opportunity to see the new classes and meet the new participants a year later.

The Power of Pomodoros

Until recently, I hadn't paid much attention to Pomodoro, though I've heard of it for a few years now. "Uncle Bob" Martin seemed to like it, and he's usually worth paying attention to in such matters. However, it mostly seemed to me like a way of organizing a variety of tasks and avoiding procrastination, and I've never had much trouble with that.

However after the January CFAR workshop suggested it in passing, I decided to give it a try; and I realized I had it all wrong. Pomodoros aren't (for me) a means of avoiding procrastination or dividing time among projects. They're a way of blasting through Ugh fields.

The Pomodoro technique is really simple compared to more involved systems like Getting Things Done (GTD). Here it is:

1. Set a timer for 25 minutes
2. Work on one thing for that 25 minutes, nothing else. No email, no phone calls, no snack breaks, no Twitter, no IM, etc.
3. Take a five minute break
4. Pick a new project, or the same project, if you prefer.
5. Repeat

That's pretty much it. You can buy a book or a special timer for this; but there's really nothing else to it. It takes longer to explain the name than the technique. (When Francesco Cirillo invented this technique in the 1980s, he was using an Italian kitchen timer shaped like a tomato. *Pomodoro* is Italian for *tomato*.)

I got interested in Pomodoro when I realized I could use it to clean my office/desk/apartment. David Allen's GTD system appealed to me, but I could never maintain it, and the 2+ days it needed to get all the way to a clean desk was always a big hurdle to vault. However, spending 25 minutes at a time, followed by a break and another project seemed a lot more manageable.

I tried it, and it worked. My desk stack quickly shrunk, not to empty, but at least to a place where an accidental elbow swing no longer launched avalanches of paper onto the floor as I typed.

So I decided to try Pomodoro on my upcoming book. The publisher was using a new authoring system and template that I was unfamiliar with. There were a dozen little details to figure out about the new system--how to check out files in git, how to create a section break, whether to use hard or soft wrapping, etc.--and I just worked through them one by one. 25 minutes later I'd knocked them all out, and was familiar enough with the new system to begin writing in earnest. I didn't know everything about the software, but I knew enough that it was no longer averting. Next I used 25 minutes on a chapter that was challenging me, and Pomodoro got me to the point where I was in the flow.

That's when I realized that Pomodoro is not a system for organizing time or avoiding procrastination (at least not for me). What it is, is an incredibly effective way to break through tasks that look too hard: code you're not familiar with, an office that's too cluttered, a chapter you don't know how to begin.

The key is that a Pomodoro forces you to focus on the unfamiliar, difficult, aversive task for 25 minutes. 25 minutes of focused attention without distractions from other, easier tasks is enough to figure out many complex situations or at least get far enough along that the next step is obvious.

Here's another example. I had a task to design a GWT widget and plug it into an existing application, and I have never done any work with GWT. Every time I looked at the frontend application code, it seemed like a big mess of confused, convoluted, dependency injected, late bound, spooky-action-at-a-distance spaghetti. Now doubtless there wasn't anything fundamentally more difficult about this code than the server side code I have been writing; and if my career had taken just a slightly different path over the last six years, frontend GWT code might be my bread and butter. But my career didn't take that path, and this code was a big Ugh field for me. So I set the Pomodoro timer on my smartphone and started working. Did I finish? No, but I got started, made progress, and proved to myself that GWT wasn't all that challenging after all. The widget is still difficult enough and GWT complex enough that I may need several more Pomodoros to finish the job, but I did get way further and learn more in 25 minutes of intense focus than I would have done in a day or even a week without it.

I don't use the Pomodoro technique exclusively. Once I get going on a project or a chapter, I don't need the help; and five minute breaks once I'm in the flow just distract me. So some days I just do 1 or 2 or 0 Pomodoros, whatever it takes to get me rolling again and past the blocker.

I also don't know if this works for genuinely difficult problems. For instance, I don't know if it will help with a difficult mathematical proof I've been struggling with for months (though I intend to find out). But for subjects that I know I can do, but can't quite figure out how to do, or where to start, the power of focusing 25 minutes of real attention on just that one problem is astonishing.

Pascal's Muggle: Infinitesimal Priors and Strong Evidence

Followup to: Pascal's Mugging: Tiny Probabilities of Vast Utilities, The Pascal's Wager Fallacy Fallacy, Being Half-Rational About Pascal's Wager Is Even Worse

Short form: Pascal's Muggle

tl;dr: If you assign superexponentially infinitesimal probability to claims of large impacts, then apparently you should ignore the possibility of a large impact even after seeing huge amounts of evidence. If a poorly-dressed street person offers to save $10^{(10\hat{}100)}$ lives (googolplex lives) for $5 using their Matrix Lord powers, and you claim to assign this scenario less than $10^{-(10\hat{}100)}$ probability, then apparently you should continue to believe absolutely that their offer is bogus even after they snap their fingers and cause a giant silhouette of themselves to appear in the sky. For the same reason, any evidence you encounter showing that the human species could create a sufficiently large number of descendants - no matter how normal the corresponding laws of physics appear to be, or how well-designed the experiments which told you about them - must be rejected out of hand. There is a possible reply to this objection using Robin Hanson's anthropic adjustment against the probability of large impacts, and in this case you will treat a Pascal's Mugger as having decision-theoretic importance exactly proportional to the Bayesian strength of evidence they present you, without quantitative dependence on the number of lives they claim to save. This however corresponds to an odd mental state which some, such as myself, would find unsatisfactory. In the end, however, I cannot see any better candidate for a prior than having a leverage penalty plus a complexity penalty on the prior probability of scenarios.

In late 2007 I coined the term "Pascal's Mugging" to describe a problem which seemed to me to arise when combining conventional decision theory and conventional epistemology in the obvious way. On conventional epistemology, the prior probability of hypotheses diminishes exponentially with their complexity; if it would take 20 bits to specify a hypothesis, then its prior probability receives a 2^{-20} penalty factor and it will require evidence with a likelihood ratio of 1,048,576:1 - evidence which we are 1048576 times more likely to see if the theory is true, than if it is false - to make us assign it around 50-50 credibility. (This isn't as hard as it sounds. Flip a coin 20 times and note down the exact sequence of heads and tails. You now believe in a state of affairs you would have assigned a million-to-one probability beforehand - namely, that the coin would produce the exact sequence HTHHHHTHTTH... or whatever - after experiencing sensory data which are more than a million times more probable if that fact is true than if it is false.) The problem is that although this kind of prior probability penalty may seem very strict at first, it's easy to construct physical scenarios that grow in size vastly faster than they grow in complexity.

I originally illustrated this using Pascal's Mugger: A poorly dressed street person says "I'm actually a Matrix Lord running this world as a computer simulation, along with many others - the universe above this one has laws of physics which allow me easy access to vast amounts of computing power. Just for fun, I'll make you an offer - you give me five dollars, and I'll use my Matrix Lord powers to save $3\uparrow\uparrow\uparrow\uparrow3$ people inside my simulations from dying and let them live long and happy lives" where $\uparrow$ is Knuth's up-arrow notation. This was originally posted in 2007, when I was a bit more naive

about what kind of mathematical notation you can throw into a random blog post without creating a stumbling block. (E.g.: On several occasions now, I've seen someone on the Internet approximate the number of dust specks from [this scenario](#) as being a "billion", since any incomprehensibly large number equals a billion.) Let's try an easier (and *way* smaller) number instead, and suppose that Pascal's Mugger offers to save a googolplex lives, where a googol is 10^{100} (a 1 followed by a hundred zeroes) and a googolplex is 10 to the googol power, so $10^{10^{100}}$ or $10^{10,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000}$ lives saved if you pay Pascal's Mugger five dollars, if the offer is honest.

If Pascal's Mugger had only offered to save a mere *googol* lives (10^{100}), we could perhaps reply that although the notion of a Matrix Lord may sound simple to say in English, if we actually try to imagine all the machinery involved, it works out to a substantial amount of computational complexity. (Similarly, Thor is a worse explanation for lightning bolts than the laws of physics because, among other points, an anthropomorphic deity is more complex than calculus in *formal* terms - it would take a larger computer program to simulate Thor as a complete mind, than to simulate Maxwell's Equations - even though in mere human words Thor sounds much easier to explain.) To imagine this scenario in formal detail, we might have to write out the laws of the higher universe the Mugger supposedly comes from, the Matrix Lord's state of mind leading them to make that offer, and so on. And so (we reply) when mere verbal English has been translated into a formal hypothesis, the Kolmogorov complexity of this hypothesis is more than 332 bits - it would take more than 332 ones and zeroes to specify - where $2^{-332} \sim 10^{-100}$. Therefore (we conclude) the net expected value of the Mugger's offer is still tiny, once its prior improbability is taken into account.

But once Pascal's Mugger offers to save a *googolplex* lives - offers us a scenario whose value is constructed by twice-repeated exponentiation - we seem to run into some difficulty using this answer. Can we really claim that the complexity of this scenario is on the order of a googol bits - that to formally write out the hypothesis would take one hundred billion billion times more bits than there are atoms in the observable universe?

And a tiny, paltry number like a googolplex is only the beginning of computationally simple numbers that are unimaginably huge. Exponentiation is defined as repeated multiplication: If you see a number like 3^5, it tells you to multiply five 3s together: $3\times3\times3\times3\times3 = 243$. Suppose we write 3^5 as $3\uparrow5$, so that a single arrow $\uparrow$ stands for exponentiation, and let the double arrow $\uparrow\uparrow$ stand for repeated exponentation, or *tetration*. Thus $3\uparrow\uparrow3$ would stand for $3\uparrow(3\uparrow3)$ or $3^{3^3} = 3^{27} = 7,625,597,484,987$. Tetration is also written as follows: ${}^{3}3 = 3\uparrow\uparrow3$. Thus ${}^{4}2 = 2^{2^{2^2}} = 2^{2^4} = 2^{16} = 65,536$. Then pentation, or repeated tetration, would be written with $3\uparrow\uparrow\uparrow3 = {}^{{}^{3}3}3 = {}^{7,625,597,484,987}3 = 3^{3^{\cdots^{3}}}$ where the ... summarizes an exponential tower of 3s seven trillion layers high.

But $3\uparrow\uparrow\uparrow3$ is still quite simple computationally - we could describe a small Turing machine which computes it - so a hypothesis involving $3\uparrow\uparrow\uparrow3$ should not therefore get a large complexity penalty, if we're penalizing hypotheses by algorithmic complexity.

I had originally intended the scenario of Pascal's Mugging to point up what seemed like a basic problem with combining conventional epistemology with conventional decision theory: Conventional epistemology says to penalize hypotheses by an exponential factor of computational complexity. This seems pretty strict in everyday life: "What? for a mere 20 bits I am to be called a million times less probable?" But for stranger hypotheses about things like Matrix Lords, the size of the hypothetical universe can blow up enormously faster than the exponential of its complexity. This would mean that all our decisions were dominated by tiny-seeming probabilities (on the order of 2^{-100} and less) of scenarios where our lightest action affected $3\uparrow\uparrow 4$ people... which would *in turn* be dominated by even *more* remote probabilities of affecting $3\uparrow\uparrow 5$ people...

This problem is worse than just giving five dollars to Pascal's Mugger - our expected utilities don't converge at all! Conventional epistemology tells us to sum over the predictions of all hypotheses weighted by their computational complexity and evidential fit. This works fine with *epistemic* probabilities and sensory predictions because no hypothesis can predict more than probability 1 or less than probability 0 for a sensory experience. As hypotheses get more and more complex, their contributed predictions have tinier and tinier weights, and the sum converges quickly. But decision theory tells us to calculate expected *utility* by summing the utility of each possible outcome, times the probability of that outcome conditional on our action. If hypothetical utilities can grow faster than hypothetical probability diminishes, the contribution of an average term in the series will keep increasing, and this sum will never converge - not if we try to do it the same way we got our epistemic predictions, by summing over complexity-weighted possibilities. (See also this similar-but-different paper by Peter de Blanc.)

Unfortunately I failed to make it clear in my original writeup that this was where the problem came from, and that it was general to situations beyond the Mugger. Nick Bostrom's writeup of Pascal's Mugging for a philosophy journal used a Mugger offering a quintillion days of happiness, where a quintillion is merely $1{,}000{,}000{,}000{,}000{,}000{,}000 = 10^{18}$. It takes at least two exponentiations to outrun a singly-exponential complexity penalty. I would be willing to assign a probability of less than 1 in 10^{18} to a random person being a Matrix Lord. You may not have to invoke $3\uparrow\uparrow\uparrow 3$ to cause problems, but you've got to use something like $10^{10^{100}}$ - double exponentiation or better. Manipulating ordinary hypotheses about the ordinary physical universe taken at face value, which just contains 10^{80} atoms within range of our telescopes, should not lead us into such difficulties.

(And then the phrase "Pascal's Mugging" got *completely* bastardized to refer to an emotional feeling of being mugged that some people apparently get when a high-stakes charitable proposition is presented to them, *regardless of whether it's supposed to have a low probability.* This is enough to make me regret having ever invented the term "Pascal's Mugging" in the first place; and for further thoughts on this see The Pascal's Wager Fallacy Fallacy (just because the stakes are high does not mean the probabilities are low, and Pascal's Wager is fallacious because of the low probability, not the high stakes!) and Being Half-Rational About Pascal's Wager Is Even Worse. Again, when dealing with issues the mere size of the apparent universe, on the order of 10^{80} - for *small* large numbers - we do *not* run into the sort of decision-theoretic problems I originally meant to single out by the concept of "Pascal's Mugging". My rough intuitive stance on x-risk charity is that if you are one of the tiny fraction of all sentient beings who happened to be born here on Earth before the intelligence explosion, when the existence of the whole vast intergalactic future

depends on what we do now, you should expect to find yourself surrounded by a *smorgasbord* of opportunities to affect small large numbers of sentient beings. There is then no reason to worry about tiny probabilities of having a large impact when we can expect to find medium-sized opportunities of having a large impact, so long as we restrict ourselves to impacts no larger than the size of the known universe.)

One proposal which has been floated for dealing with Pascal's Mugger in the decision-theoretic sense is to penalize hypotheses that let you affect a large number of people, in proportion to the number of people affected - what we could call perhaps a "leverage penalty" instead of a "complexity penalty".

Unfortunately this potentially leads us into a different problem, that of *Pascal's Muggle.*

Suppose a poorly-dressed street person asks you for five dollars in exchange for doing a googolplex's worth of good using his Matrix Lord powers.

"Well," you reply, "I think it very improbable that I would be able to affect so many people through my own, personal actions - who am I to have such a great impact upon events? Indeed, I think the probability is somewhere around one over googolplex, maybe a bit less. So no, I won't pay five dollars - it is unthinkably improbable that I could do so much good!"

"I see," says the Mugger.

A wind begins to blow about the alley, whipping the Mugger's loose clothes about him as they shift from ill-fitting shirt and jeans into robes of infinite blackness, within whose depths tiny galaxies and stranger things seem to twinkle. In the sky above, a gap edged by blue fire opens with a horrendous tearing sound - you can hear people on the nearby street yelling in sudden shock and terror, implying that they can see it too - and displays the image of the Mugger himself, wearing the same robes that now adorn his body, seated before a keyboard and a monitor.

"That's not actually me," the Mugger says, "just a conceptual representation, but I don't want to drive you insane. Now give me those five dollars, and I'll save a googolplex lives, just as promised. It's easy enough for me, given the computing power my home universe offers. As for why I'm doing this, there's an ancient debate in philosophy among my people - something about how we ought to sum our expected utilities - and I mean to use the video of this event to make a point at the next decision theory conference I attend. Now will you give me the five dollars, or not?"

"Mm... no," you reply.

"*No?*" says the Mugger. "I understood earlier when you didn't want to give a random street person five dollars based on a wild story with no evidence behind it. But now I've offered you evidence."

"Unfortunately, you haven't offered me *enough* evidence," you explain.

"Really?" says the Mugger. "I've opened up a fiery portal in the sky, and that's not enough to persuade you? What do I have to do, then? Rearrange the planets in your solar system, and wait for the observatories to confirm the fact? I suppose I could also explain the true laws of physics in the higher universe in more detail, and let you play around a bit with the computer program that encodes all the universes containing the googolplex people I would save if you gave me the five dollars -"

"Sorry," you say, shaking your head firmly, "there's just no *way* you can convince me that I'm in a position to affect a googolplex people, because the prior probability of that is one over googolplex. If you wanted to convince me of some fact of merely 2^{-100} prior probability, a mere decillion to one - like that a coin would come up heads and tails in some particular pattern of a hundred coinflips - then you could just show me 100 bits of evidence, which is within easy reach of my brain's sensory bandwidth. I mean, you could just flip the coin a hundred times, and my eyes, which send my brain a hundred megabits a second or so - though that gets processed down to one megabit or so by the time it goes through the lateral geniculate nucleus - would easily give me enough data to conclude that this decillion-to-one possibility was true. But to conclude something whose prior probability is on the order of one over googolplex, I need on the order of a googol bits of evidence, and you can't present me with a sensory experience containing a googol bits. Indeed, you can't *ever* present a mortal like me with evidence that has a likelihood ratio of a googolplex to one - evidence I'm a googolplex times more likely to encounter if the hypothesis is true, than if it's false - because the chance of all my neurons spontaneously rearranging themselves to fake the same evidence would always be higher than one over googolplex. You know the old saying about how once you assign something probability one, or probability zero, you can never change your mind regardless of what evidence you see? Well, odds of a googolplex to one, or one to a googolplex, work pretty much the same way."

"So no matter what evidence I show you," the Mugger says - as the blue fire goes on crackling in the torn sky above, and screams and desperate prayers continue from the street beyond - "you can't ever notice that you're in a position to help a googolplex people."

"Right!" you say. "I can believe that you're a Matrix Lord. I mean, I'm not a *total* Muggle, I'm psychologically capable of responding in *some* fashion to that giant hole in the sky. But it's just completely forbidden for me to assign any significant probability whatsoever that you will actually save a googolplex people after I give you five dollars. You're lying, and I am absolutely, absolutely, absolutely confident of that."

"So you weren't *just* invoking the leverage penalty as a plausible-sounding way of getting out of paying me the five dollars earlier," the Mugger says thoughtfully. "I mean, I'd understand if that was just a rationalization of your discomfort at forking over five dollars for what seemed like a tiny probability, when I hadn't done my duty to present you with a corresponding amount of evidence before demanding payment. But you... you're acting like an AI would if it was actually programmed with a leverage penalty on hypotheses!"

"Exactly," you say. "I'm forbidden *a priori* to believe I can ever do that much good."

"Why?" the Mugger says curiously. "I mean, all I have to do is press this button here and a googolplex lives will be saved." The figure within the blazing portal above points to a green button on the console before it.

"Like I said," you explain again, "the prior probability is just too infinitesimal for the massive evidence you're showing me to overcome it -"

The Mugger shrugs, and vanishes in a puff of purple mist.

The portal in the sky above closes, taking with the console and the green button.

(The screams go on from the street outside.)

A few days later, you're sitting in your office at the physics institute where you work, when one of your colleagues bursts in through your door, seeming highly excited. "I've got it!" she cries. "I've figured out that whole dark energy thing! Look, these simple equations retrodict it exactly, there's no way that could be a coincidence!"

At first you're also excited, but as you pore over the equations, your face configures itself into a frown. "No..." you say slowly. "These equations may look extremely simple so far as computational complexity goes - and they do exactly fit the petabytes of evidence our telescopes have gathered so far - but I'm afraid they're far too improbable to ever believe."

"What?" she says. "Why?"

"Well," you say reasonably, "if these equations are actually true, then our descendants will be able to exploit dark energy to do computations, and according to my back-of-the-envelope calculations here, we'd be able to create around a googolplex people that way. But that would mean that we, here on Earth, are in a position to affect a googolplex people - since, if we blow ourselves up via a nanotechnological war or *(cough)* make certain other errors, those googolplex people will never come into existence. The prior probability of us being in a position to impact a googolplex people is on the order of one over googolplex, so your equations must be wrong."

"Hmm..." she says. "I hadn't thought of that. But what if these equations are right, and yet somehow, everything I do is exactly balanced, down to the googolth decimal point or so, with respect to how it impacts the chance of modern-day Earth participating in a chain of events that leads to creating an intergalactic civilization?"

"How would *that* work?" you say. "There's only seven billion people on today's Earth - there's probably been only a hundred billion people who ever existed total, or will exist before we go through the intelligence explosion or whatever - so even before analyzing your exact position, it seems like your leverage on future affairs couldn't reasonably be less than a one in ten trillion part of the future or so."

"But then given this physical theory which seems obviously true, my acts might imply expected utility differentials on the order of $10^{10^{100}-13}$," she explains, "and I'm not allowed to believe that no matter how much evidence you show me."

This problem may not be as bad as it looks; with some further reasoning, the leverage penalty may lead to more sensible behavior than depicted above.

Robin Hanson has suggested that the logic of a leverage penalty should stem from the general improbability of individuals being in a *unique* position to affect many others (which is why I called it a leverage penalty). At most 10 out of $3\uparrow\uparrow\uparrow 3$ people can ever be in a position to be "solely responsible" for the fate of $3\uparrow\uparrow\uparrow 3$ people if "solely responsible" is taken to imply a causal chain that goes through no more than 10 people's decisions; i.e. at most 10 people can ever be solely_{10} responsible for any given event. Or if "fate" is taken to be a sufficiently ultimate fate that there's at most 10 other decisions of similar magnitude that could cumulate to determine someone's outcome utility to within ±50%, then any given person could have their fate_{10} determined on at most 10 occasions. We would surely agree, while assigning priors at the dawn of reasoning, that an agent randomly selected from the pool of all

agents in Reality has at most a 100/X chance of being able to be solely$_{10}$ responsible for the fate$_{10}$ of X people. Any reasoning we do about universes, their complexity, sensory experiences, and so on, should maintain this net balance. You can even strip out the part about agents and carry out the reasoning on pure causal nodes; the chance of a randomly selected causal node being in a unique$_{100}$ position on a causal graph with respect to $3\uparrow\uparrow\uparrow 3$ other nodes ought to be at most $100/3\uparrow\uparrow\uparrow 3$ for finite causal graphs. (As for infinite causal graphs, well, if problems arise *only* when introducing infinity, maybe it's infinity that has the problem.)

Suppose we apply the Hansonian leverage penalty to the face-value scenario of our own universe, in which there are apparently no aliens and the galaxies we can reach in the future contain on the order of 10^{80} atoms; which, if the intelligence explosion goes well, might be transformed into on the very loose order of... let's ignore a lot of intermediate calculations and just call it the equivalent of 10^{80} centuries of life. (The neurons in your brain perform lots of operations; you don't get only one computing operation per element, because you're powered by the Sun over time. The universe contains a lot more negentropy than just 10^{80} bits due to things like the gravitational potential energy that can be extracted from mass. Plus we should take into account reversible computing. But of course it also takes more than one computing operation to implement a century of life. So I'm just going to xerox the number 10^{80} for use in these calculations, since it's not supposed to be the main focus.)

Wouldn't it be terribly odd to find ourselves - where by 'ourselves' I mean the hundred billion humans who have ever lived on Earth, for no more than a century or so apiece - solely$_{100,000,000,000}$ responsible for the fate$_{10}$ of around 10^{80} units of life? Isn't the prior probability of this somewhere around 10^{-68}?

Yes, according to the leverage penalty. But a prior probability of 10^{-68} is not an insurmountable epistemological barrier. If you're taking things at face value, 10^{-68} is just 226 bits of evidence or thereabouts, and your eyes are sending you a megabit per second. Becoming convinced that *you,* yes *you* are an Earthling is epistemically doable; you just need to see a stream of sensory experiences which is 10^{68} times more probable if you are an Earthling than if you are someone else. If we take everything at face value, then there could be around 10^{80} centuries of life over the history of the universe, and only 10^{11} of those centuries will be lived by creatures who discover themselves occupying organic bodies. Taking everything at face value, the sensory experiences of your life are unique to Earthlings and should immediately convince you that you're an Earthling - just looking around the room you occupy will provide you with sensory experiences that plausibly belong to only 10^{11} out of 10^{80} life-centuries.

If we *don't* take everything at face value, then there might be such things as ancestor simulations, and it might be that your experience of looking around the room is something that happens in 10^{20} ancestor simulations for every time that it happens in 'base level' reality. In this case your probable leverage on the future is diluted (though it may be large even post-dilution). But this is not something that the Hansonian leverage penalty *forces* you to believe - not when the putative stakes are still as small as 10^{80}. Conceptually, the Hansonian leverage penalty doesn't interact much with the Simulation Hypothesis (SH) at all. If you don't believe SH, then you think that the experiences of creatures like yours are rare in the universe and hence present strong, convincing evidence for you occupying the leverage-privileged position of an Earthling - much stronger evidence than its prior improbability. (There's some separate

anthropic issues here about whether or not this is *itself* evidence for SH, but I don't think that question is intrinsic to leverage penalties per se.)

A key point here is that even if you accept a Hanson-style leverage penalty, it doesn't have to manifest as an inescapable commandment of modesty. You need not refuse to believe (in your deep and irrevocable humility) that you could be someone as special as an Ancient Earthling. Even if Earthlings matter in the universe - even if we occupy a unique position to affect the future of galaxies - it is still possible to encounter pretty convincing evidence that you're an Earthling. Universes the size of 10^{80} do not pose problems to conventional decision-theoretic reasoning, or to conventional epistemology.

Things play out similarly if - still taking everything at face value - you're wondering about the chance that you could be special even for an Earthling, because you might be one of say 10^4 people in the history of the universe who contribute a major amount to an x-risk reduction project which ends up actually saving the galaxies. The vast majority of the improbability here is just in being an Earthling in the first place! Thus most of the clever arguments for not taking this high-impact possibility at face value would also tell you not to take being an Earthling at face value, since Earthlings as a whole are much more unique within the total temporal history of the universe than you are supposing yourself to be unique among Earthlings. But given ¬SH, the prior improbability of being an Earthling can be overcome by a few megabits of sensory experience from looking around the room and querying your memories - it's not like 10^{80} is enough future beings that the number of agents randomly hallucinating similar experiences outweighs the number of real Earthlings. Similarly, if you don't think lots of Earthlings are hallucinating the experience of going to a donation page and clicking on the Paypal button for an x-risk charity, that sensory experience can easily serve to distinguish you as one of 10^4 people donating to an x-risk philanthropy.

Yes, there are various clever-sounding lines of argument which involve *not* taking things at face value - "Ah, but maybe you should consider yourself as an indistinguishable part of this here large reference class of deluded people who think they're important." Which I consider to be a bad idea because it renders you a permanent Muggle by putting you into an inescapable reference class of self-deluded people and then dismissing all your further thoughts as insufficient evidence because you *could* just be deluding yourself further about whether these are good arguments. Nor do I believe the world can only be saved by good people who are incapable of distinguishing themselves from a large class of crackpots, all of whom have no choice but to continue based on the tiny probability that they are not crackpots. (For more on this see Being Half-Rational About Pascal's Wager Is Even Worse.) In this case you are a Pascal's Muggle not because you've explicitly assigned a probability like one over googolplex, but because you took an improbability like 10^{-6} at unquestioning face value and then cleverly questioned all the evidence which could've overcome that prior improbability, and so, in practice, you can never climb out of the epistemological sinkhole. By the same token, you should conclude that you are just self-deluded about being an Earthling since real Earthlings are so rare and privileged in their leverage.

In general, leverage penalties don't translate into advice about modesty or that you're just deluding yourself - they just say that to be rationally coherent, your picture of the universe has to imply that your sensory experiences are at least as rare as the corresponding magnitude of your leverage.

Which brings us back to Pascal's Mugger, in the original alleyway version. The Hansonian leverage penalty seems to imply that to be coherent, *either* you believe that your sensory experiences are *really actually* 1 in a googolplex - that only 1 in a googolplex beings experiences what you're experiencing - or else you really *can't* take the situation at face value.

Suppose the Mugger is telling the truth, and a googolplex other people are being simulated. Then there are at least a googolplex people in the universe. Perhaps some of them are hallucinating a situation similar to this one by sheer chance? Rather than telling you flatly that you can't have a large impact, the Hansonian leverage penalty implies a coherence requirement on how uniquely you think your sensory experiences identify the position you believe yourself to occupy. When it comes to believing you're one of 10^{11} Earthlings who can impact 10^{80} other life-centuries, you need to think your sensory experiences are unique to Earthlings - identify Earthlings with a likelihood ratio on the order of 10^{69}. This is quite achievable, if we take the evidence at face value. But when it comes to improbability on the order of $1/3\uparrow\uparrow\uparrow 3$, the prior improbability *is* inescapable - your sensory experiences *can't* possibly be that unique - which is assumed to be appropriate because almost-everyone who ever believes they'll be in a position to help $3\uparrow\uparrow\uparrow 3$ people *will in fact* be hallucinating. Boltzmann brains should be much more common than people in a unique position to affect $3\uparrow\uparrow\uparrow 3$ others, at least if the causal graphs are finite.

Furthermore - although I didn't realize this part until recently - applying Bayesian updates from that starting point may partially avert the Pascal's Muggle effect:

Mugger: "Give me five dollars, and I'll save $3\uparrow\uparrow\uparrow 3$ lives using my Matrix Powers."

You: "Nope."

Mugger: "Why not? It's a really large impact."

You: "Yes, and I assign a probability on the order of 1 in $3\uparrow\uparrow\uparrow 3$ that I would be in a unique position to affect $3\uparrow\uparrow\uparrow 3$ people."

Mugger: "Oh, is that really the probability that you assign? Behold!"

(A gap opens in the sky, edged with blue fire.)

Mugger: "Now what do you think, eh?"

You: "Well... I can't actually say this observation has a likelihood ratio of $3\uparrow\uparrow\uparrow 3$ to 1. No stream of evidence that can enter a human brain over the course of a century is ever going to have a likelihood ratio larger than, say, $10^{10^{26}}$ to 1 at the *absurdly most,* assuming one megabit per second of sensory data, for a century, each bit of which has at least a 1-in-a-trillion error probability. I'd probably start to be dominated by Boltzmann brains or other exotic minds well before then."

Mugger: "So you're not convinced."

You: "Indeed not. The probability that you're telling the truth is so tiny that God couldn't find it with an electron microscope. Here's the five dollars."

Mugger: "Done! You've saved $3\uparrow\uparrow\uparrow 3$ lives! Congratulations, you're never going to top that, your peak life accomplishment will now always lie in your past. But why'd you give me the five dollars if you think I'm lying?"

You: "Well, because the evidence you *did* present me with had a likelihood ratio of at least a billion to one - I would've assigned less than 10^{-9} prior probability of seeing this when I woke up this morning - so in accordance with Bayes's Theorem I promoted the probability from 1/3↑↑↑3 to at least 10^9/3↑↑↑3, which when multiplied by an impact of 3↑↑↑3, yields an expected value of at least a billion lives saved for giving you five dollars."

I confess that I find this line of reasoning a bit suspicious - it seems overly clever. But on the level of intuitive virtues of rationality, it does seem less stupid than the original Pascal's Muggle; this muggee is at least *behaviorally* reacting to the evidence. In fact, they're reacting in a way exactly proportional to the evidence - they would've assigned the same net importance to handing over the five dollars if the Mugger had offered 3↑↑↑4 lives, so long as the strength of the evidence seemed the same.

(Anyone who tries to apply the lessons here to actual x-risk reduction charities (which I think is probably a bad idea), keep in mind that the vast majority of the improbable-position-of-leverage in any x-risk reduction effort comes from being an Earthling in a position to affect the future of a hundred billion galaxies, and that sensory evidence for being an Earthling is what gives you most of your belief that your actions can have an outsized impact.)

So why not just run with this - why not just declare the decision-theoretic problem resolved, if we have a rule that seems to give reasonable behavioral answers in practice? Why not just go ahead and program that rule into an AI?

Well... I still feel a bit nervous about the idea that Pascal's Muggee, after the sky splits open, is handing over five dollars while claiming to assign probability on the order of 10^9/3↑↑↑3 that it's doing any good.

I think that my own reaction in a similar situation would be along these lines instead:

Mugger: "Give me five dollars, and I'll save 3↑↑↑3 lives using my Matrix Powers."

Me: "Nope."

Mugger: "So then, you think the probability I'm telling the truth is on the order of 1/3↑↑↑3?"

Me: "Yeah... that probably *has* to follow. I don't see any way around that revealed belief, given that I'm not actually giving you the five dollars. I've heard some people try to claim silly things like, the probability that you're telling the truth is counterbalanced by the probability that you'll kill 3↑↑↑3 people instead, or something else with a conveniently equal and opposite utility. But there's no way that things would balance out *exactly* in practice, if there was no *a priori* mathematical requirement that they balance. Even if the prior probability of your saving 3↑↑↑3 people and killing 3↑↑↑3 people, conditional on my giving you five dollars, *exactly* balanced down to the log(3↑↑↑3) decimal place, the likelihood ratio for your telling me that you would "save" 3↑↑↑3 people would not be exactly 1:1 for the two hypotheses down to the log(3↑↑↑3) decimal place. So if I assigned probabilities much greater than 1/3↑↑↑3 to your doing something that affected 3↑↑↑3 people, my actions would be overwhelmingly dominated by even a tiny difference in likelihood ratio elevating the probability that you saved 3↑↑↑3 people over the probability that

you did something bad to them. The only way this hypothesis can't dominate my actions - really, the only way my expected utility sums can converge at all - is if I assign probability on the order of 1/3↑↑↑3 or less. I don't see any way of escaping that part."

Mugger: "But can you, in your mortal uncertainty, truly assign a probability as low as 1 in 3↑↑↑3 to any proposition whatever? Can you truly believe, with your error-prone neural brain, that you could make 3↑↑↑3 statements *of any kind* one after another, and be wrong, on average, about once?"

Me: "Nope."

Mugger: "So give me five dollars!"

Me: "Nope."

Mugger: "Why not?"

Me: "Because even though I, in my mortal uncertainty, will eventually be wrong about all sorts of things if I make enough statements one after another, this fact can't be used to increase the probability of arbitrary statements beyond what my prior says they should be, because then my prior would sum to more than 1. There must be some kind of required condition for taking a hypothesis seriously enough to worry that I might be overconfident about it -"

Mugger: "Then behold!"

(A gap opens in the sky, edged with blue fire.)

Mugger: "Now what do you think, eh?"

Me *(staring up at the sky):* "...whoa." *(Pause.)* "You turned into a cat."

Mugger: "What?"

Me: "Private joke. Okay, I think I'm going to have to rethink a *lot* of things. But if you want to tell me about how I was wrong to assign a prior probability on the order of 1/3↑↑↑3 to your scenario, I will shut up and listen very carefully to what you have to say about it. Oh, and here's the five dollars, can I pay an extra twenty and make some other requests?"

(The thought bubble pops, and we return to two people standing in an alley, the sky above perfectly normal.)

Mugger: "Now, in this scenario we've just imagined, you were taking my case seriously, right? But the evidence there couldn't have had a likelihood ratio of more than $10^{10^{26}}$ to 1, and probably much less. So by the method of imaginary updates, you must assign probability at least $10^{-10^{26}}$ to my scenario, which when multiplied by a benefit on the order of 3↑↑↑3, yields an unimaginable bonanza in exchange for just five dollars -"

Me: "Nope."

Mugger: "How can you possibly say that? You're not being logically coherent!"

Me: "I agree that I'm not being logically coherent, but I think that's acceptable in this case."

Mugger: "This ought to be good. Since when are rationalists allowed to deliberately be logically incoherent?"

Me: "Since we don't have infinite computing power -"

Mugger: "That sounds like a fully general excuse if I ever heard one."

Me: "No, this is a *specific* consequence of bounded computing power. Let me start with a simpler example. Suppose I believe in a set of mathematical axioms. Since I don't have infinite computing power, I won't be able to know all the deductive consequences of those axioms. And *that* means I will necessarily fall prey to the conjunction fallacy, in the sense that you'll present me with a theorem X that is a deductive consequence of my axioms, but which I don't know to be a deductive consequence of my axioms, and you'll ask me to assign a probability to X, and I'll assign it 50% probability or something. Then you present me with a brilliant lemma Y, which clearly seems like a likely consequence of my mathematical axioms, and which also seems to imply X - once I see Y, the connection from my axioms to X, via Y, becomes obvious. So I assign P(X&Y) = 90%, or something like that. Well, that's the conjunction fallacy - I assigned P(X&Y) > P(X). The thing is, if you *then* ask me P(X), after I've seen Y, I'll reply that P(X) is 91% or at any rate something higher than P(X&Y). I'll have changed my mind about what my prior beliefs logically imply, because I'm not logically omniscient, even if that looks like assigning probabilities *over time* which are incoherent in the Bayesian sense."

Mugger: "And how does this work out to my not getting five dollars?"

Me: "In the scenario you're asking me to imagine, you present me with evidence which I currently think Just Plain Shouldn't Happen. And if that actually *does* happen, the sensible way for me to react is by questioning my prior assumptions and the reasoning which led me assign such low probability. One way that I handle my lack of logical omniscience - my finite, error-prone reasoning capabilities - is by being willing to assign infinitesimal probabilities to non-privileged hypotheses so that my prior over all possibilities can sum to 1. But if I actually see strong evidence for something I previously thought was super-improbable, I don't just do a Bayesian update, I should also question whether I was right to assign such a tiny probability in the first place - whether it was really as complex, or unnatural, as I thought. In real life, you are not ever supposed to have a prior improbability of 10^{-100} for some fact distinguished enough to be written down in advance, and yet encounter strong evidence, say 10^{10} to 1, that the thing has actually happened. If something like that happens, you don't do a Bayesian update to a posterior of 10^{-90}. Instead you question both whether the evidence might be weaker than it seems, *and* whether your estimate of prior improbability might have been poorly calibrated, because rational agents who actually have well-calibrated priors should not encounter situations like that until they are ten billion days old. Now, this may mean that I end up doing some non-Bayesian updates: I say some hypothesis has a prior probability of a quadrillion to one, you show me evidence with a likelihood ratio of a billion to one, and I say 'Guess I was wrong about that quadrillion to one thing' rather than being a Muggle about it. And then I shut up and listen to what *you* have to say about how to estimate probabilities, because on my worldview, I wasn't *expecting* to see you turn into a cat. But for me to make a super-update like that - reflecting a posterior belief that I was logically incorrect about the prior probability - you have to really actually show me the

evidence, you can't just ask me to imagine it. This is something that only logically incoherent agents ever say, but that's all right because I'm not logically omniscient."

At some point, we're going to have to build some sort of actual prior into, you know, some sort of actual self-improving AI.

(Scary thought, right?)

So far as I can presently see, the logic requiring some sort of leverage penalty - not just so that we don't pay $5 to Pascal's Mugger, but also so that our expected utility sums converge at all - seems clear enough that I can't yet see a good alternative to it (feel welcome to suggest one), and Robin Hanson's rationale is by far the best I've heard.

In fact, what we actually need is more like a combined leverage-and-complexity penalty, to avoid scenarios like this:

Mugger: "Give me $5 and I'll save 3↑↑↑3 people."

You: "I assign probability exactly 1/3↑↑↑3 to that."

Mugger: "So that's one life saved for $5, on average. That's a pretty good bargain, right?"

You: "Not by comparison with x-risk reduction charities. But I also like to do good on a smaller scale now and then. How about a penny? Would you be willing to save 3↑↑↑3/500 lives for a penny?"

Mugger: "Eh, fine."

You: "Well, the probability of that is 500/3↑↑↑3, so here's a penny!" *(Goes on way, whistling cheerfully.)*

Adding a complexity penalty *and* a leverage penalty is necessary, not just to avert this exact scenario, but so that we don't get an infinite expected utility sum over a 1/3↑↑↑3 probability of saving 3↑↑↑3 lives, 1/(3↑↑↑3 + 1) probability of saving 3↑↑↑3 + 1 lives, and so on. If we combine the standard complexity penalty with a leverage penalty, the whole thing should converge.

Probability penalties are epistemic features - they affect what we believe, not just what we do. Maps, ideally, correspond to territories. Is there any territory that this complexity+leverage penalty can correspond to - any state of a single reality which would make these the true frequencies? Or is it only interpretable as pure uncertainty over realities, with there being no single reality that could correspond to it? To put it another way, the complexity penalty and the leverage penalty seem unrelated, so perhaps they're mutually inconsistent; can we show that the union of these two theories has a model?

As near as I can figure, the corresponding state of affairs to a complexity+leverage prior improbability would be a Tegmark Level IV multiverse in which each reality got an amount of magical-reality-fluid corresponding to the complexity of its program (1/2 to the power of its Kolmogorov complexity) and then this magical-reality-fluid had to

be *divided* among all the causal elements within that universe - if you contain $3\uparrow\uparrow\uparrow 3$ causal nodes, then each node can only get $1/3\uparrow\uparrow\uparrow 3$ of the total realness of that universe. (As always, the term "magical reality fluid" reflects an attempt to demarcate a philosophical area where I feel quite confused, and try to use correspondingly blatantly wrong terminology so that I do not mistake my reasoning about my confusion for a solution.) This setup is not entirely implausible because the Born probabilities in our own universe look like they might behave like this sort of magical-reality-fluid - quantum amplitude flowing between configurations in a way that preserves the total amount of realness while dividing it between worlds - and perhaps every other part of the multiverse must necessarily work the same way for some reason. It seems worth noting that part of what's motivating this version of the 'territory' is that our sum over all real things, weighted by reality-fluid, can then converge. In other words, the reason why complexity+leverage works in decision theory is that the union of the two theories has a model in which the total multiverse contains an amount of reality-fluid that can sum to 1 rather than being infinite. (Though we need to suppose that either (a) only programs with a finite number of causal nodes exist, or (2) programs can divide finite reality-fluid among an infinite number of nodes via some measure that gives every experience-moment a well-defined relative amount of reality-fluid. Again see caveats about basic philosophical confusion - perhaps our map needs this property over its uncertainty but the territory doesn't have to work the same way, etcetera.)

If an AI's overall architecture is also such as to enable it to carry out the "You turned into a cat" effect - where if the AI actually ends up with strong evidence for a scenario it assigned super-exponential improbability, the AI reconsiders its priors and the apparent strength of evidence rather than executing a blind Bayesian update, though this part is formally a tad underspecified - then at the moment I can't think of anything else to add in.

In other words: This is my best current idea for how a prior, e.g. as used in an AI, could yield decision-theoretic convergence over explosively large possible worlds.

However, I would still call this a semi-open FAI problem (edit: wide-open) because it seems quite plausible that somebody is going to kick holes in the overall view I've just presented, or come up with a better solution, possibly within an hour of my posting this - the proposal is both recent and weak even by my standards. I'm also worried about whether it turns out to imply anything crazy on anthropic problems. Over to you, readers.

Robustness of Cost-Effectiveness Estimates and Philanthropy

Note: I formerly worked as a research analyst at GiveWell. This post describes the evolution of my thinking about robustness of cost-effectiveness estimates in philanthropy. All views expressed here are my own.

Up until 2012, I believed that detailed explicit cost-effectiveness estimates are very important in the context of philanthropy. My position was reflected in a comment that I made in 2011:

> The problem with using unquantified heuristics and intuitions is that the "true" expected values of philanthropic efforts plausibly differ by many orders of magnitude, and unquantified heuristics and intuitions are frequently insensitive to this. The last order of magnitude is the only one that matters; all others are negligible by comparison. So if at all possible, one should do one's best to pin down the philanthropic efforts with the "true" expected value per dollar of the highest (positive) order of magnitude. It seems to me as though any feasible strategy for attacking this problem involves explicit computation.

During my time at GiveWell, my position on this matter shifted. I still believe that there are instances in which *rough* cost-effectiveness estimates can be useful for determining good philanthropic foci. But I've shifted toward the position that **effective altruists should spend much more time on qualitative analysis than on quantitative analysis in determining how they can maximize their positive social impact**.

In this post I'll focus on one reason for my shift: **explicit cost-effectiveness estimates are generally much less robust than I had previously thought**.

The history of GiveWell's estimates for lives saved per dollar

Historically, GiveWell used "cost per life saved" as a measure of the cost-effectiveness of its global health recommendations. Examination of the trajectory of GiveWell's cost-effectiveness estimates shows that **GiveWell has consistently updated in the direction of its ranked charities having higher "cost per life saved" than GiveWell had previously thought.** I give the details below.

The discussion should be read with the understanding that **donating to GiveWell's top charities has benefits that extend beyond saving lives**, so that "number of lives saved" understates cost-effectiveness..

At the end of each of 2009 and 2010, GiveWell named VillageReach its #1 ranked charity. VillageReach estimated the cost-per-life-saved of its pilot project as being < $200, and at the end of 2009, GiveWell gave a "conservative" estimate of $545/life saved. In 2011, GiveWell reassessed VillageReach's pilot project, commending VillageReach for being transparent enough for reassessment to be possible, and concluding that

> We feel that within the framework of "delivering proven, cost-effective interventions to improve health," AMF and SCI are solidly better giving opportunities than VillageReach (both now and at the time when we recommended it). Given the information we have, we see less room for doubt in the cases for AMF's and SCI's impact than in the case for VillageReach's.

Here "AMF" refers to Against Malaria Foundation, which is GiveWell's current #1 ranked charity. If AMF is currently more cost-effective than VillageReach was at the time when GiveWell recommended VillageReach, then the best cost-per-life-saved figure for GiveWell's recommended charities is (and was) the cost-effectiveness of donating to AMF.

AMF delivers long-lasting insecticide treated nets (LLINs) to the developing world to protect people against mosquitoes that spread malaria. This contrasts with VillageReach, which works to increase vaccination rates. Vaccines are thought to be more cost-effective than LLINs, and GiveWell has not been able to find strong giving opportunities in vaccination, so the cost per life saved of the best opportunity that GiveWell has found for individual donors is correspondingly higher.

At the end of 2011, GiveWell estimated that the marginal cost per life associated with donating to AMF at $1600/life saved. During 2012, I vetted GiveWell's page on LLINs and uncovered an issue, which led GiveWell to revise its estimate for AMF's marginal cost per life saved to $2300/life saved at the end of 2012. This does not take into account regression to the mean, which can be expected to raise the cost per life saved.

The discussion above shows a consistent trend in the direction of the marginal cost per life saved in the developing world being higher than initially meets the eye. Note that the difference between VillageReach's original estimate and GiveWell's current estimate is about an order of magnitude.

Concrete factors that further reduce the expected value of donating to AMF

A key point that I had missed when I thought about these things earlier in my life is that **there are many small probability failure modes which are not significant individually, but which collectively substantially reduce cost-effectiveness**. When I encountered such a potential failure mode, my reaction was to think "this is very unlikely to be an issue" and then to forget about it. I didn't notice that I was doing this many times in a row.

I list many relevant factors that reduce AMF's expected cost-effectiveness below. Some of these are from GiveWell's discussion of possible negative or offsetting impacts in GiveWell's review of AMF. Others are implicitly present in GiveWell's review of AMF and GiveWell's review of LLINs, and others are issues that have emerged in the interim. I would emphasize that **I don't think that any of the points listed is a big issue** and that **GiveWell and AMF take precautionary efforts to guard against them**. But I think that they *collectively* reduce cost-effectiveness by a substantial amount.

- If GiveWell's customers weren't funding AMF, another funder might, and that funder might instead be funding much less effective activities.
- If AMF weren't working in a given region, there might be other organizations that would deliver LLINs to that region, and these other organizations may instead be

funding much less effective activities.
- It could be that the workers who distribute the LLINs would otherwise be providing more cost-effective health care interventions.
- The five RCTs that found that LLIN distribution reduces mortality could be systematically flawed in a non-obvious way.
- While the Cochrane Review that contains a meta-analysis of the RCTs referred to unpublished studies so as to counteract publication bias, there may be unpublished studies that were missed, and which were not published, because they found no effect.
- The field workers who are assigned to distribute LLINs may steal the nets to sell them for a profit.
- Fathers may steal nets from pregnant mothers and sell them for a profit.
- LLIN recipients may use the nets for fishing.
- LLIN users may not fasten LLINs properly.
- Mosquitoes may develop biological resistance to the insecticide used on LLINs.
- Mosquitoes may develop "behavioral resistance" to the insecticides used on LLINs by evolving to bite during the day (when LLINs are not used) rather than during the night.

Most of the relevant factors will vary by region where AMF ships nets, and some may be present in certain locations and not others.

Do these considerations argue against donating to AMF?

In view of the issues above, one might wonder whether it's better to donate to a charity in a different cause, or better not to donate at all. Some relevant points follow:

Donating to AMF has benefits beyond saving lives. The above discussion of cost-effectiveness figures concerns "cost per life saved" specifically. But there are benefits to donating to AMF that go beyond saving lives.

- Malaria control reduces the morbidity of malaria. A Cochrane Review of the health benefits of LLINs reports on reductions in anemia, enlarged spleen, and other health outcomes.
- People are more productive when they're healthy than they are when they're ill.
- There is some evidence that malaria control increases children's income later on in life.
- The above benefits could be massively leveraged via flow-through effects.

Updates in the direction of reduced cost-effectiveness aren't specific to global health. Based on my experience at GiveWell, I've found that *regardless* of the cause within which one investigates giving opportunities, there's a strong tendency for giving opportunities to appear progressively less promising as one learns more. AMF and LLIN distribution have stood up to scrutiny *unusually well*. It remains the case that Global health and nutrition may be an unusually good cause for individual donors.

Updates in the direction of reduced LLIN cost-effectiveness push in favor of cash transfers over LLINs. Transferring cash to people in the developing world is an unusually straightforward intervention. While there are potential downsides to transferring cash, there seem to be fewer potential failure modes associated with it than there are potential failure modes associated with LLIN distribution. There are strong arguments that favor LLINs over cash transfers, but difference in straightforwardness of the interventions in juxtaposition with the phenomenon of

surprisingly large updates in the direction of reduced cost-effectiveness is a countervailing consideration.

Why do cost-effectiveness updates skew so negatively?

When I first started thinking seriously about philanthropy in 2009, I thought that if one has impressions of a philanthropic opportunity, one will be equally likely to update in the direction of it being better than meets the eye as one will be to update the direction of the opportunity being worse than meets the eye. So I was surprised to discover how strong the tendency is for philanthropic opportunities to look worse over time rather than better over time.

Aside from the empirical data, something that shifted my view is Holden's observation that outlier cost-effectiveness estimates need to be regressed to one's Bayesian prior over the values of all possible philanthropic opportunities. Another reason for my shift is GiveWell finding that philanthropic markets are more efficient than it had previously thought. I think that optimism bias also plays a role.

This is all consistent with GiveWell's view that one should expect good giving to be hard.

Implications for maximizing cost-effectiveness

The remarks and observations above imply that **Bayesian regression in the context of philanthropy is substantially larger than expected**. This favors:

- Examining a philanthropic opportunity from many angles rather than relying too heavily on a single perspective.
- Giving more weight to robust inputs into one's assessment of a philanthropic opportunity. Estimating the cost-effectiveness of health interventions in the developing world has proved to be exceedingly difficult, and this pushes in favor of giving more weight to inputs for which it's possible to make relatively well-grounded assessments. Some of these are room for more funding, the quality of the people behind a project and historical precedent.
- Choosing giving opportunities that it will be possible to learn from, and giving now instead of giving later when one encounters such an opportunity.
- Choosing giving opportunities about which one has a lot of information. GiveWell has been moving away from the old criterion of recommending proven interventions, and giving more weight to upside relative to track record than GiveWell used to. However, this partially reflects the discovery that the expected effectiveness of ostensibly "proven" interventions is lower than previously thought.

10-Step Anti-Procrastination Checklist

Despite recent strides in my productivity habits, I still catch myself procrastinating at work more often than I'd like. It's not that I make a conscious decision to put off a project; it just feels as though I wake up 20 minutes later and realize that nothing got accomplished. (Or, to avoid the passive voice and take much-deserved responsibility, I "realize that I haven't accomplished anything".)

I've been looking for techniques to improve, and got a lot out of LukeProg's articles on How to Beat Procrastination and My Algorithm for Beating Procrastination, based on Piers Steel's The Procrastination Equation.

But I also wanted a way to put the principles to use with the lowest activation cost possible. I can't expect unmotivated future-me to be too cooperative; I need to provide him with an easy path to get in flow.

So! I developed a 10-Step Productivity Checklist, pulling the concepts from Luke's articles and adding a couple points that are important for me. Now whenever I notice myself being unproductive I have a much easier time following the steps one by one until I get back in a good mindset to work.

Productivity Checklist:

1. **What is the task?** Make sure you're going to focus on one thing at a time.
2. **Do you have something to drink?** Get yourself some tea, coffee, or water.
3. **Are distractions closed?** Shut the door, quit Tweetdeck, close the Facebook and Gmail tabs, and set skype to "Do not disturb."
4. **What music will you listen to inspire yourself to be productive or get in flow?** Put on a good instrumental playlist! (I love video game soundtracks, further notes in comments.)
5. **Why are you doing this task?** Trace the value until you feel the benefit.
6. **What are the parts to this task?** Break things down as much as you can, until they're physical actions if possible.
7. **What are some ways to gamify the task?** Try to have fun with it!
8. **What are some rewards you can offer yourself for completing sections of the task?** Smiling, throwing your arms up in the air and proclaiming victory, or M&M's all count.
9. **What's an achievable goal for this sitting?** Set a reasonable expectation for yourself.
10. **How long will you work until you take a break?** Set a timer and commit to focusing.

Get into flow!

I'd love to hear from you:

- Whether these are useful
- Any ideas for good ways to enact these steps
- Steps that should be added/removed/tweaked
- Whether there are other posts/resources that you've found valuable

I hope this helps you as much as it's helping me, and that together we can make it even better!

LW Study Hall - 2 Month Update

Comment reposted from (link) for exposure

Two months have passed and I'm glad to say the LW Study Hall on tinychat is still active and alive. Since judging from the comments it kind of looks like we've moved on from tinychat, a review like this might be useful for anyone who hasn't been there yet.

My first sessions on the chat were driven more by curiosity than anything else since I didn't believe it would be really effective for me – I've felt that I procrastinate too much, but it never occurred to me that working together with other people might make me more effective. I was proven wrong.

Since those first sessions I've been online almost every day and got to see different people come and go, and some people stay. It didn't take long for me to feel like a part of the "chat community", and to feel motivated to work to see the regulars more often, some of which I might even consider friends now. The atmosphere is friendly, people make an active effort to integrate newcomers in the "community" and I have yet to see an argument that isn't constructive. Though the breaks are a bit flexible, people usually don't overstretch it and it's generally good practice not to chat during a working phase. More introverted people can participate without taking part in the chat much and without broadcasting video.

So, what makes this chat so effective in combating procrastination? Pomodoros are the "flow" of the chat. Since you're working with other people, you are much more likely to stick to the pomodoro cycle than if you set those constraints for yourself. That doesn't just mean you keep the breaks relatively short, but you also don't work too long. I find that if I work alone, I tend to keep at it for longer than I can keep concentrated. When I do take a break I don't really have anything else to do, so I might start to procrastinate, leading to a work cycle where the "breaks" can be as long as the working phases. This has been my main issue with structuring my working day, and I was more surprised than I probably should have been to see that problem solved by working in a group. Judging from my own experiences and those of others I believe everyone struggling with akrasia should at least try if it works for him/her. For those who struggle with akrasia more, it might be useful to combine several techniques such as precommitting to fixed working dates, showing your screen on camera or finding someone on the chat who will remind you (e.g. via skype) to show up again if you've been absent for longer (or any number of other methods like beeminder).

There are a few issues with the chat, especially that tinychat isn't always stable. The limited options have also been subject of complaints, but it's so far the best thing we've found. I'm optimistic that a better option will be found or created in the long term – the more people frequent the chat, the more likely it gets. Covering all time slots hasn't worked out perfectly, but we usually have good "coverage" during the UTC afternoon/evening, so that is probably a good time to try. In case the chat is empty, don't be discouraged, just try again later. I will try to put as many of my working hours in the precommitment schedule (link on top of the chat window) and hope others will do so more often too, so it's possible to sync up working time.

Over these two months the lesswrong chat has become a substantial part of my life that I really want to keep, ideally for much longer. While it is no longer an experiment for me, I want to invite you to try it, if you haven’t already. I’d be glad to welcome you on the chat anytime. :)

Wikifying the blog list

Konkvistador's excellent List of Blogs by LWers led me to some of my favorite blogs, but is pretty well hidden and gradually becoming obsolete. In order to create an easily-update-able replacement, I have created the wiki page List of Blogs and added most of the blogs from Konkvistador's list. If you have a blog, or you read blogs, please help in the following ways:

-- Add your blog if it's not on there, and *if it has updated in the past few months* (no dead blogs this time, exceptions for very complete archives of excellent material like Common Sense Atheism in the last section)

-- Add any other blogs you like that are written by LWers or frequently engage with LW ideas

-- Remove your blog if you don't want it on there (I added some prominent critics of LW ideas who might not want to be linked to us)

-- Move your blog to a different category if you don't like the one it's in right now

-- Add a description of your blog, or change the one that already exists

-- Change the name you're listed by (I defaulted to people's LW handles)

-- Bold the name of your blog if it updates near-daily, has a large readership/commentership, and/or gets linked to on LW a lot

-- Improve formatting

Somebody more familiar with the Less Wrong twittersphere might want to do something similar to Grognor's Less Wrong on Twitter

Be Nice to Non-Rationalists

Note: I have no intention of criticizing the person involved. I admire that (s)he made the "right" decision in the end (in my opinion), and I mention it only as an example we could all learn from. I did request permission to use his/her anecdote here. I'll also use the pronoun "he" when really I mean he/she.

Once Pat says “no,” it’s harder to get to “yes” than if you had never asked.

Crocker's rules has this very clear clause, and we should keep it well in mind:

Note that Crocker's Rules does not mean you can insult people; it means that other people don't have to worry about whether they are insulting you. Crocker's Rules are a discipline, not a privilege. Furthermore, taking advantage of Crocker's Rules does not imply reciprocity. How could it? Crocker's Rules are something you do for yourself, to maximize information received - not something you grit your teeth over and do as a favor.

Recently, a rationalist heard over social media that an acquaintance - a friend-of-a-friend - had found their lost pet. They said it was better than winning a lottery. The rationalist responded that unless they'd spent thousands of dollars searching, or posted a large reward, then they're saying something they don't really mean. Then, feeling like a party-pooper and a downer, he deleted his comment.

I believe this was absolutely the correct things to do. As Miss Manners says (http://www.washingtonpost.com/wp-dyn/content/article/2007/02/06/AR2007020601518.html), people will associate unpleasant emotions with the source and the cause. They're not going to say, oh, that's correct; I was mistaken about the value of my pet; thank you for correcting my flawed value system.

Instead they'll say, those rationalists are so heartless, attaching dollar signs to everything. They think they know better. They're rude and stuck up. I don't want to have anything to do with them. And then they'll think walk away with a bad impression of us. (Yes, all of us, for we are a minority now, and each of us reflects upon all of us, the same way a Muslim bomber would reflect poorly on public opinion of all Muslims, while a Christian bomber would not.) In the future they'll be less likely to listen to any one of us.

The only appropriate thing to say in this case is "I'm so happy for you." But that doesn't mean we can't promote ourselves ever. Here are some alternatives.

- At another time, ask for "help" with your own decisions. Go through the process of calculating out all the value and expected values. This is completely non-confrontational, and your friends/acquaintances will not need to defend anything. Whenever they give a suggestion, praise it as being a good idea, and then make a show of weighing the expected value out loud.
- Say "wow, I don't know many people who'd spend that much! Your pet is lucky to have someone like you!" But it must be done without any sarcasm. They might

feel a bit uncomfortable taking that much praise. They might go home and mull it over.
- Invite them to "try something you saw online" with you. This thing could be mindcharting, the estimation game, learning quantum physics, meditation, goal refactoring, anything. Emphasize the curiosity/exploring aspect. See if it leads into a conversation about rationality. Don't mention the incident with the pet - it could come off as criticism.
- At a later date, introduce them to Methods or Rationality. Say it's because "it's funny," or "you have a lot of interesting ideas," or even just "I think you'll like it." That's generally a good starting point. :)
- Let it be. First do no harm.

I was told long ago (in regards to LGBT rights) that minds are not changed by logic or reasoning or facts. They are changed over a long period of time by emotions. For us, that means showing what we believe without pressing it on others, while at the same time being the kind of person you want to be like. If we are successful and happy, if we carry ourselves with kindness and dignity, we'll win over hearts.

What do professional philosophers believe, and why?

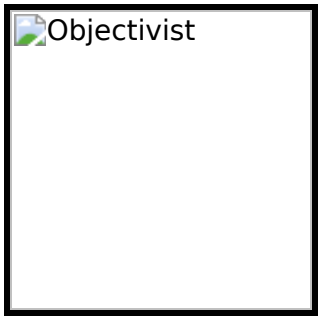
LessWrong has twice discussed the **PhilPapers Survey** of professional philosophers' views on thirty controversies in their fields — in early 2011 and, more intensively, in **late 2012**. We've also been having some lively debates, **prompted by LukeProg**, about the general value of contemporary philosophical assumptions and methods. It would be swell to test some of our intuitions about how philosophers go wrong (and right) by looking closely at the aggregate output and conduct of philosophers, but relevant data is hard to come by.

Fortunately, Davids Chalmers and Bourget have done a lot of the work for us. They released a **paper summarizing the PhilPapers Survey results** two days ago, identifying, by factor analysis, seven major components consolidating correlations between philosophical positions, influences, areas of expertise, etc.

1. **Anti-Naturalists**: Philosophers of this stripe tend (more strongly than most) to assert libertarian free will (correlation with factor .66), theism (.63), the metaphysical possibility of zombies (.47), and A theories of time (.28), and to reject physicalism (.63), naturalism (.57), personal identity reductionism (.48), and liberal egalitarianism (.32).

Anti-Naturalists tend to work in philosophy of religion (.3) or Greek philosophy (.11). They avoid philosophy of mind (-.17) and cognitive science (-.18) like the plague. They hate Hume (-.14), Lewis (-.13), Quine (-.12), analytic philosophy (-.14), and being from Australasia (-.11). They love Plato (.13), Aristotle (.12), and Leibniz (.1).

2. **Objectivists**: They tend to accept 'objective' moral values (.72), aesthetic values (.66), abstract objects (.38), laws of nature (.28), and scientific posits (.28). Note 'Objectivism' is being used here to pick out a tendency to treat value as objectively binding and metaphysical posits as objectively real; it isn't connected to Ayn Rand.



A disproportionate number of objectivists work in normative ethics (.12), Greek philosophy (.1), or philosophy of religion (.1). They don't work in philosophy of science (-.13) or biology (-.13), and aren't continentalists (-.12) or Europeans (-.14). Their favorite philosopher is Plato (.1), least favorites Hume (-.2) and Carnap (-.12).

3. **Rationalists**: They tend to self-identify as 'rationalists' (.57) and 'non-naturalists' (.33), to accept that some knowledge is *a priori* (.79), and to assert that some truths

are analytic, i.e., 'true by definition' or 'true in virtue of 'meaning' (.72). Also tend to posit metaphysical laws of nature (.34) and abstracta (.28). 'Rationalist' here clearly isn't being used in the LW or freethought sense; philosophical rationalists as a whole in fact tend to be theists.

Rationalists are wont to work in metaphysics (.14), and to avoid thinking about the sciences of life (-.14) or cognition (-.1). They are extremely male (.15), inordinately British (.12), and prize Frege (.18) and Kant (.12). They absolutely despise Quine (-.28, the largest correlation for a philosopher), and aren't fond of Hume (-.12) or Mill (-.11) either.

4. **Anti-Realists**: They tend to define truth in terms of our cognitive and epistemic faculties (.65) and to reject scientific realism (.6), a mind-independent and knowable external world (.53), metaphysical laws of nature (.43), and the notion that proper names have no meaning beyond their referent (.35).

They are extremely female (.17) and young (.15 correlation coefficient for year of birth). They work in ethics (.16), social/political philosophy (.16), and 17th-19th century philosophy (.11), avoiding metaphysics (-.2) and the philosophies of mind (-.15) and language (-.14). Their heroes are Kant (.23), Rawls (.14), and, interestingly, Hume (.11). They avoid analytic philosophy even more than the anti-naturalists do (-.17), and aren't fond of Russell (-.11).

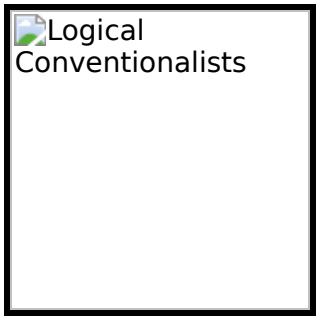
5. **Externalists**: Really, they just like everything that anyone calls 'externalism'. They think the content of our mental lives in general (.66) and perception in particular (.55), and the justification for our beliefs (.64), all depend significantly on the world outside our heads. They also think that you can fully understand a moral imperative without being at all motivated to obey it (.5).

Beyond externalism, they really have very little in common. They avoid 17th-18th century philosophy (-.13), and tend to be young (.1) and work in the UK (.1), but don't converge upon a common philosophical tradition or area of expertise, as far as the survey questions indicated.

6. ***Star Trek* Haters**: This group is less clearly defined than the above ones. The main thing uniting them is that they're thoroughly convinced that teleportation would mean death (.69). Beyond that, Trekophobes tend to be deontologists (.52) who don't switch on trolley dilemmas (.47) and like A theories of time (.41).

Trekophobes are relatively old (-.1) and American (.13 affiliation). They are quite rare in Australia and Asia (-.18 affiliation). They're fairly evenly distributed across philosophical fields, and tend to avoid weirdo intuitions-violating naturalists — Lewis (-.13), Hume (-.12), analytic philosophers generally (-.11).

7. **Logical Conventionalists**: They two-box on Newcomb's Problem (.58), reject nonclassical logics (.48), and reject epistemic relativism and contextualism (.48). So they love causal decision theory, think all propositions/facts are generally well-behaved (always either true or false and never both or neither), and think there are always facts about which things you know, independent of who's evaluating you. Suspiciously normal.



They're also fond of a wide variety of relatively uncontroversial, middle-of-the-road views most philosophers agree about or treat as 'the default' — political egalitarianism (.33), abstract object realism (.3), and atheism (.27). They tend to think zombies are metaphysically possible (.26) and to reject personal identity reductionism (.26) — which aren't metaphysically innocent or uncontroversial positions, but, again, do seem to be remarkably straightforward and banal approaches to all these problems. Notice that a lot of these positions are intuitive and 'obvious' in isolation, but that they don't converge upon any coherent world-view or consistent methodology. They clearly aren't hard-nosed philosophical conservatives like the Anti-Naturalists, Objectivists, Rationalists, and Trekophobes, but they also clearly aren't upstart radicals like the Externalists (on the analytic side) or the Anti-Realists (on the continental side). They're just kind of, well... obvious.

Conventionalists are the only identified group that are strongly analytic in orientation (.19). They tend to work in epistemology (.16) or philosophy of language (.12), and are rarely found in 17th-19th century (-.12) or continental (-.11) philosophy. They're influenced by notorious two-boxer and modal realist David Lewis (.1), and show an aversion to Hegel (-.12), Aristotle (-.11), and and Wittgenstein (-.1).

An observation: Different philosophers rely on — and fall victim to — substantially different groups of methods and intuitions. A few simple heuristics, like 'don't believe weird things until someone conclusively demonstrates them' and 'believe things that seem to be important metaphysical correlates for basic human institutions' and 'fall in love with any views starting with "ext"', explain a surprising amount of diversity. And there are clear common tendencies to either **trust one's own rationality** or to distrust it in partial (Externalism) or pathological (Anti-Realism, Anti-Naturalism) ways. But the heuristics don't hang together in a single Philosophical World-View or Way Of Doing Things, or even in two or three such world-views.

There is no large, coherent, consolidated group that's particularly attractive to LWers across the board, but philosophers seem to fall short of LW expectations for some quite distinct reasons. So attempting to criticize, persuade, shame, praise, or even *speak of or address* philosophers as a whole may be a bad idea. I'd expect it to be more productive to target specific 'load-bearing' doctrines on dimensions like the above than to treat the group as a monolith, for many of the same reasons we don't want to treat 'scientists' or 'mathematicians' as monoliths.

Another important result: Something is going seriously wrong with the *high-level* **training** and **enculturation** of professional philosophers. Or fields are just attracting thinkers who are disproportionately bad at critically assessing a number of the basic claims their field is predicated on or exists to assess.

Philosophers working in decision theory are drastically *worse* at Newcomb than are other philosophers, two-boxing 70.38% of the time where non-specialists two-box 59.07% of the time (normalized after getting rid of 'Other' answers). Philosophers of religion are the most likely to get questions about religion wrong — 79.13% are theists (compared to 13.22% of non-specialists), and they tend strongly toward the Anti-Naturalism dimension. Non-aestheticians think aesthetic value is objective 53.64% of the time; aestheticians think it's objective 73.88% of the time. Working in epistemology tends to make you an internalist, philosophy of science tends to make you a Humean, metaphysics a Platonist, ethics a deontologist. This isn't always the case; but it's genuinely troubling to see non-expertise emerge as a predictor of getting *any* important question in an academic field right.

EDIT: I've replaced "cluster" talk above with "dimension" talk. I had in mind gjm's "*clusters in philosophical idea-space*", not distinct groups of *philosophers*. gjm makes this especially clear:

> The claim about these positions being made by the authors of the paper is not, not even a little bit, "most philosophers fall into one of these seven categories". It is "you can generally tell most of what there is to know about a philosopher's opinions if you know how well they fit or don't fit each of these seven categories". Not "philosopher-space is mostly made up of these seven pieces" but "philosopher-space is approximately seven-dimensional".

I'm particularly guilty of promoting this misunderstanding (including in portions of my own brain) by not noting that the dimensions can be flipped to speak of (anti-anti-)naturalists, anti-rationalists, etc. My apologies. As Douglas_Knight notes below, "If there are clusters [of philosophers], PCA might find them, but PCA might tell you something interesting even if there are no clusters. But if there are clusters, the factors that PCA finds won't be the clusters, but the differences between them. [...] Actually, factor analysis pretty much assumes that there aren't clusters. If factor 1 put you in a cluster, that would tell pretty much all there is to say and would pin down your factor 2, but the idea in factor analysis is that your factor 2 is designed to be as free as possible, despite knowing factor 1."