Best of LessWrong: March 2013

1. Reflection in Probabilistic Logic
2. MetaMed: Evidence-Based Healthcare
3. Recent updates to gwern.net (2012-2013)
4. Boring Advice Repository
5. Harry Potter and the Methods of Rationality Bookshelves
6. Bayesian Adjustment Does Not Defeat Existential Risk Charity
7. Schelling Day: A Rationalist Holiday
8. Rationality Habits I Learned at the CFAR Workshop
9. We Don't Drink Vodka (LW Moscow report)
10. Suggestion: Read Paul Graham
11. Don't Get Offended
12. Programming the LW Study Hall

Best of LessWrong: March 2013

1. Reflection in Probabilistic Logic
2. MetaMed: Evidence-Based Healthcare
3. Recent updates to gwern.net (2012-2013)
4. Boring Advice Repository
5. Harry Potter and the Methods of Rationality Bookshelves
6. Bayesian Adjustment Does Not Defeat Existential Risk Charity
7. Schelling Day: A Rationalist Holiday
8. Rationality Habits I Learned at the CFAR Workshop
9. We Don't Drink Vodka (LW Moscow report)
10. Suggestion: Read Paul Graham
11. Don't Get Offended
12. Programming the LW Study Hall

Reflection in Probabilistic Logic

Paul Christiano has devised **a new fundamental approach** to the "Löb Problem" wherein Löb's Theorem seems to pose an obstacle to AIs building successor AIs, or adopting successor versions of their own code, that trust the same amount of mathematics as the original. (I am currently writing up a more thorough description of the *question* this preliminary technical report is working on answering. For now the main online description is in a quick Summit talk I gave. See also Benja Fallenstein's description of the problem in the course of presenting a different angle of attack. Roughly the problem is that mathematical systems can only prove the soundness of, aka 'trust', weaker mathematical systems. If you try to write out an exact description of how AIs would build their successors or successor versions of their code in the most obvious way, it looks like the mathematical strength of the proof system would tend to be stepped down each time, which is undesirable.)

Paul Christiano's approach is inspired by the idea that whereof one cannot prove or disprove, thereof one must assign probabilities: and that although no mathematical system can contain its own *truth* predicate, a mathematical system might be able to contain a reflectively consistent *probability* predicate. In particular, it looks like we can have:

$\forall a, b: (a < P(\varphi) < b) \quad \Rightarrow \quad P(a < P('\varphi') < b) = 1$
$\forall a, b: P(a \leq P('\varphi') \leq b) > 0 \quad \Rightarrow \quad a \leq P(\varphi) \leq b$

Suppose I present you with the human and probabilistic version of a Gödel sentence, the Whitely sentence "You assign this statement a probability less than 30%." If you disbelieve this statement, it is true. If you believe it, it is false. If you assign 30% probability to it, it is false. If you assign 29% probability to it, it is true.

Paul's approach resolves this problem by restricting your belief about your own probability assignment to within epsilon of 30% for any epsilon. So Paul's approach replies, "Well, I assign *almost* exactly 30% probability to that statement - maybe a little more, maybe a little less - in fact I think there's about a 30% chance that I'm a tiny bit under 0.3 probability and a 70% chance that I'm a tiny bit over 0.3 probability." A standard fixed-point theorem then implies that a consistent assignment like this should exist. If asked if the probability is over 0.2999 or under 0.30001 you will reply with a definite yes.

We haven't yet worked out a walkthrough showing if/how this solves the Löb obstacle to self-modification, and the probabilistic theory itself is nonconstructive (we've shown that something like this should exist, but not how to compute it). Even so, a possible fundamental triumph over Tarski's theorem on the undefinability of truth and a number of standard Gödelian limitations is important news as math *qua* math, though work here is still in very preliminary stages. There are even whispers of unrestricted comprehension in a probabilistic version of set theory with $\forall \varphi: \exists S: P(x \in S) = P(\varphi(x))$, though this part is not in the preliminary report and is at even earlier stages and could easily not work out at all.

It seems important to remark on how this result was developed: Paul Christiano showed up with the idea (of consistent probabilistic reflection via a fixed-point theorem) to a week-long "math squad" (aka MIRI Workshop) with Marcello Herreshoff, Mihaly Barasz, and myself; then we all spent the next week proving that version after

version of Paul's idea couldn't work or wouldn't yield self-modifying AI; until finally, a day after the workshop was supposed to end, it produced something that looked like it might work. If we hadn't been trying to *solve* this problem (with hope stemming from how it seemed like the sort of thing a reflective rational agent ought to be able to do somehow), this would be just another batch of impossibility results in the math literature. I remark on this because it may help demonstrate that Friendly AI is a productive approach to math *qua* math, which may aid some mathematician in becoming interested.

I further note that this does not mean the Löbian obstacle is resolved and no further work is required. Before we can conclude that we need a computably specified version of the theory plus a walkthrough for a self-modifying agent using it.

See also the blog post on the MIRI site (and subscribe to MIRI's newsletter here to keep abreast of research updates).

This LW post is the preferred place for feedback on the paper.

EDIT: But see discussion on a Google+ post by John Baez here. Also see here for how to display math LaTeX in comments.

MetaMed: Evidence-Based Healthcare

In a world where 85% of doctors can't solve simple Bayesian word problems...

In a world where only 20.9% of reported results that a pharmaceutical company tries to investigate for development purposes, fully replicate...

In a world where "p-values" are anything the author wants them to be...

...and where there are all sorts of amazing technologies and techniques which nobody at your hospital has ever heard of...

...there's also **MetaMed**. Instead of just having “evidence-based medicine” in journals that doctors don't actually read, MetaMed will provide you with actual evidence-based healthcare. Their Chairman and CTO is Jaan Tallinn (cofounder of Skype, major funder of xrisk-related endeavors), one of their major VCs is Peter Thiel (major funder of MIRI), their management includes some names LWers will find familiar, and their researchers know math and stats and in many cases have also read LessWrong. If you have a sufficiently serious problem and can afford their service, MetaMed will (a) put someone on reading the relevant research literature who understands real statistics and can tell whether the paper is trustworthy; and (b) refer you to a cooperative doctor in their network who can carry out the therapies they find.

MetaMed was partially inspired by the case of a woman who had her fingertip chopped off, was told by the hospital that she was screwed, and then read through an awful lot of literature on her own until she found someone working on an advanced regenerative therapy that let her actually grow the fingertip back. The idea behind MetaMed isn't just that they will scour the literature to find how the best experimentally supported treatment differs from the average wisdom - people who regularly read LW will be aware that this is often a pretty large divergence - but that they will also look for this sort of very recent technology that most hospitals won't have heard about.

This is a new service and it has to interact with the existing medical system, so they are currently expensive, starting at $5,000 for a research report. (Keeping in mind that a basic report involves a lot of work by people who must be good at math.) If you have a sick friend who can afford it - especially if the regular system is failing them, and they want (or you want) their next step to be *more* science instead of "alternative medicine" or whatever - please do refer them to MetaMed immediately. We can’t all have nice things like this someday unless somebody pays for it while it’s still new and expensive. And the regular healthcare system really is bad enough at science (especially in the US, but science is difficult everywhere) that there's no point in condemning anyone to it when they can afford better.

I also got my hands on a copy of MetaMed's standard list of citations that they use to support points to reporters. What follows isn't nearly everything on MetaMed's list, just the items I found most interesting.

90% of preclinical cancer studies could not be replicated:
http://www.nature.com/nature/journal/v483/n7391/full/483531a.html

"It is frequently stated that it takes an average of 17 years for research evidence to reach clinical practice. Balas and Bohen, Grant, and Wratschko all estimated a time lag of 17 years measuring different points of the process." - http://www.jrsm.rsmjournals.com/content/104/12/510.full

"The authors estimated the volume of medical literature potentially relevant to primary care published in a month and the time required for physicians trained in medical epidemiology to evaluate it for updating a clinical knowledgebase.... Average time per article was 2.89 minutes, if this outlier was excluded. Extrapolating this estimate to 7,287 articles per month, this effort would require 627.5 hours per month, or about 29 hours per weekday."

One-third of hospital patients are harmed by their stay in the hospital, and 7% of patients are either permanently harmed or die: http://www.ama-assn.org/amednews/2011/04/18/prl20418.htm

(I emailed MetaMed to ask for the actual bibliography for the following citations, since that wasn't included in the copy of the list I saw. I already recognize some of the citations having to do with Bayesian reasoning, which makes me fairly confident of the others.)

Statistical Illiteracy

Doctors often confuse sensitivity and specificity (Gigerenzer 2002); most physicians do not understand how to compute the positive predictive value of a test (Hoffrage and Gigerenzer 1998); a third overestimate benefits if they are expressed as positive risk reductions (Gigerenzer et al 2007).
Physicians think a procedure is more effective if the benefits are described as a relative risk reduction rather than as an absolute risk reduction (Naylor et al 1992).
Only 3 out of 140 reviewers of four breast cancer screening proposals noticed that all four were identical proposals with the risks represented differently (Fahey et al 1995).
60% of gynecologists do not understand what the sensitivity and specificity of a test are (Gigerenzer at al 2007).
95% of physicians overestimated the probability of breast cancer given a positive mammogram by an order of magnitude (Eddy 1982).
When physicians receive prostate cancer screening information in terms of five-year survival rates, 78% think screening is effective; when the same information is given in terms of mortality rates, 5% believe it is effective (Wegwarth et al, submitted).
Only one out of 21 obstetricians could estimate the probability that an unborn child had Down syndrome given a positive test (Bramwell, West, and Salmon 2006).
Sixteen out of twenty HIV counselors said that there was no such thing as a false positive HIV test (Gigerenzer et all 1998).
Only 3% of questions in the certification exam for the American Board of Internal Medicine cover clinical epidemiology or medical statistics, and risk communication is not addressed (Gigerenzer et al 2007).
British GPs rarely change their prescribing patterns and when they do it’s rarely in response to evidence (Armstrong et al 1996).

Drug Advertising

Direct-to-customer advertising by pharmaceutical companies, which is intended to sell drugs rather than to educate, often does not contain information about a drug's success rate (only 9% did), alternative methods of treatment (29%), behavioral changes (24%), or the treatment duration (9%) (Bell et al 2000).

Patients are more likely to request advertised drugs and doctors to prescribe them, regardless of their misgivings (Gilbody et al 2005).

Medical Errors

44,000 to 98,000 patients are killed in US hospitals each year by documented, preventable medical errors (Kohn et al 2000).
Despite proven effectiveness of simple checklists in reducing infections in hospitals (Provonost et al 2006), most ICU physicians do not use them.
Simple diagnostic tools which may even ignore some data give measurably better outcomes in areas such as deciding whether to put a new admission in a coronary care bed (Green and Mehr 1997).
Tort law often actively penalizes physicians who practice evidence-based medicine instead of the medicine that is customary in their area (Monahan 2007).
Out of 175 law schools, only one requires a basic course in statistics or research methods (Faigman 1999), so many judges, jurors, and lawyers are misled by nontransparent statistics.
93% of surgeons, obstreticians, and other health care professionals at high risk for malpractice suits report practicing defensive medicine (Studdert et al 2005).

Regional Variations in Health Care

Tonsillectomies vary twelvefold between the counties in Vermont with the highest and lowest rates of the procedure (Wennberg and Gittelsohn 1973).
Fivefold variations in one-year survival from cancer across different regions have been observed (Quam and Smith 2005).
Fiftyfold variations in people receiving drug treatment for dementia has been reported (Prescribing Observatory for Mental Health 2007).
Rates of certain surgical procedures vary tenfold to fifteenfold between regions (McPherson et al 1982).
Clinicians are more likely to consult their colleagues than medical journals or the library, partially explaining regional differences (Shaughnessy et al 1994).

Research

Researchers may report only favorable trials, only report favorable data (Angell 2004), or cherry-pick data to only report favorable variables or subgroups (Rennie 1997).
Of 50 systematic reviews and meta-analyses on asthma treatment 40 had serious or extensive flaws, including all 6 associated with industry (Jadad et al 2000).
Less high-tech knowledge and applications tend to be considered less innovative and ignored (Shi and Singh 2008).

Poor Use of Statistics In Research

Only about 7% of major-journal trials report results using transparent statistics (Nuovo, Melnivov and Chang 2002).
Data are often reported in biased ways: for instance, benefits are often reported as relative risks (“reduces the risk by half”) and harms as absolute risks (“an increase of 5 in 1000”); absolute risks seem smaller even when the risk is the same (Gigerenzer et al 2007).
Half of trials inappropriately use significance tests for baseline comparison; 2/3 present subgroup findings, a sign of possible data fishing, often without appropriate tests for interaction (Assman et al 2000).

One third of studies use mismatched framing, where benefits are reported one way (usually relative risk reduction, which makes them look bigger) and harms another (usually absolute risk reduction, which makes them look smaller) (Sedrakyan and Shih 2007).

Positive Publication Bias

Positive publication bias overstates the effects of treatment by up to one-third (Schultz et al 1995).
More than 50% of research is unpublished or unreported (Mathieu et al 2009).
In ten high-impact medical journals, only 45.5% of trials were adequately registered before testing began; of these 31% show discrepancies between outcomes measured and published (Mathieu et al 2009).

Pharmaceutical Company Induced Bias

Studies funded by the pharmaceutical industry are more likely to report results favorable to the sponsoring company (Lexchin et al 2003).
There is a significant association between industry sponsorship and both pro-industry outcomes and poor methodology (Bekelman and Kronmal 2008).
In manufacturer-supported trials of non-steroidal anti-inflammatory drugs, half the time the data presented did not match claims made within the article (Rochon et al 1994).
68% of US health research is funded by industry (Research!America 2008), which means that research that leads to profits to the health care industry tends to be prioritized.
71 out of 78 drugs approved by the FDA in 2002 are “me too” drugs that are more profitable because of the patent but not substantially different from existing medication (Angell 2004).
“Seeding trials” by pharmaceutical companies promote treatments instead of testing hypotheses (Hill et al 2008).
Even accurate research may be misreported by pharmaceutical company advertising, including ads in medical journals (Villanueva et al 2003).
In 92% of cases, pharmaceutical leaflets distributed to doctors have data summaries that either cannot be verified or inaccurately summarize available data (Kaiser et al 2004).

I don't plan on becoming seriously sick, but if I do, I think I'll check in with MetaMed just to make sure nobody is ignoring the research results showing that you shouldn't feed the patient rat poison.

Recent updates to gwern.net (2012-2013)

Previous: Recent updates to gwern.net (2011)

"But where shall wisdom be found? / And where is the place of understanding? / Man knoweth not the price thereof; neither is it found in the land of the living…for the price of wisdom is above rubies."

As before, here is material I've worked on in the 477 days since my last update which LWers may find interesting. In roughly chronological & topical order, here are the major additions to `gwern.net`:

- I interviewed translator Michael House about his work in Japan as a translator
- finished data collection for my hafu anime statistics page and begun analysis. (I've achieved good coverage of characters, found an astonishingly consistent absence of Korean characters, and confirmed the blond-haired/blue-eyed stereotype; but my original thesis doesn't seem to work and the data is too unevenly distributed to identify time trends.)
- judged the 2011 & 2012 results for the Haskell Summer of Codes and the accuracy of my predictions
- did a meta-analysis on whether dual n-back increases IQ, and examining possible biases and various claims about what makes the training work or not work
- did another meta-analysis on whether iodine increases IQ, etc

- modafinil:
    - checked for subjective effects of blinded modafinil
    - updated my modafinil price-chart twice, and expanded with brand data and a new armodafinil table
    - researched modafinil-related prosecutions & convictions in the USA
    - and any connection with schizophrenia
- tried kratom
- did a nicotine gum/n-back experiment
- did 2 potassium experiments; neither improved my mood/productivity, and one damaged my sleep
- my Silk Road page has been expanded with a BBC interview, putting SR in a historical cypherpunk context, an updated account of all arrests & law enforcement actions, and application of basic statistics to ordering
- ran 2 sleep experiments on the timing of taking a vitamin D supplement: I found that taking vitamin D before bed substantially damaged my sleep, while taking vitamin D after waking up did not hurt & somewhat helped
- checked whether a walking desk (treadmill) damaged typing speed or accuracy
- I have run 3 Wikipedia experiments establishing that: Talk page edits are ignored by editors; random link deletions (and their restoration) are also ignored by editors; and external link suggestions on Talk pages are also ignored by readers. (I take the former 2 as indicative of the decline in edit activity and rise of deletionist beliefs on Wikipedia.)
- tried some economic/historical analysis: "Reasons of State: Why Didn't Denmark Sell Greenland to the USA?"

- Defending sunk costs essay (LW discussion)
- "Slowing Moore's Law: Why You Might Want To and How You Would Do It"
- "The Hyperbolic Time Chamber as Brain Emulation Analogy"
- tried estimating the bandwidth of a Death Note
- compiled predictions for *Harry Potter and the Methods of Rationality*
- looked into Conscientiousness and online education; studies so far are useless from a meta-analytic standpoint
- tripled length of appendix dealing with the reliability of mainstream science (methodological flaws, replication rates, etc)
- finished meta-ethics essay, "The Narrowing Circle"
- explained the philosophy saying "one man's modus ponens is another man's modus tollens"
- speculation about a restoration of the British monarchy
- clean up & exploratory data analysis of SDr's lucid dreaming data
- Who wrote the *Death Note* script? (LW discussion)
- 2012 US election predictions: statistical comparison

- Comment anchoring experiment (LW discussion)

- Turing-completeness in surprising places (inventory of particularly "weird machines"; relevant to computer and AI security)

Transcribed or translated:

- Nash's letters on cryptography
- Douglas Hofstadter's superrationality columns (from *Metamagical Themas*, 1985)
- "The Iron Law Of Evaluation And Other Metallic Rules", Rossi 1987 (lessons from the large RCTs evaluating social & welfare interventions)
- "The Ups and Downs of the Hope Function In a Fruitless Search", Falk et al 1994
- Gene Wolfe on writing
- "Shiny balls of Mud: William Gibson Looks at Japanese Pursuits of Perfection" (2002)
- "Otaku Talk", Okada et al 2004
- "Earth in My Window", Murakami 2005
- "On The Battlefield of 'Superflat'"
- "Ero-Anime: Manga Comes Alive", Sarrazin 2010
- 1996 *NewType* interview with Hideaki Anno (translated by me, with the help of an EGFer)
- 1997 *Animeland* interview with Hideaki Anno (bought, transcribed, and translated by me with the help of other LWers)
- 1997 *Utena* interviews

More technical:

- added edit history statistics/visualization for `gwern.net` using GitStats
- site traffic updates: July-December 2011, January 2012-July 2012, July 2012-Jan 2013
- There's also been a lot of backend changes: switching to Amazon S3+Cloudflare, adding error pages, metadata like tags, A/B testing, but no need to go into detail.

Personal:

- dumped my notes on my 2011 visit to San Francisco

- posted summaries of my personality & attitudes & my RSS feed collection
- enjoyed some mead; I still like tea better, though
- dumped notes on the 2012 SF convention ICON

Boring Advice Repository

This is an extension of a comment I made that I can't find and also a request for examples. It seems plausible that, when giving advice, many people optimize for deepness or punchiness of the advice rather than for actual practical value. There may be good reasons to do this - e.g. advice that sounds deep or punchy might be more likely to be listened to - but as a corollary, there could be valuable advice that people generally don't give because it doesn't sound deep or punchy. Let's call this **boring advice**.

An example that's been discussed on LW several times is "make checklists." Checklists are great. We should totally make checklists. But "make checklists" is not a deep or punchy thing to say. Other examples include "google things" and "exercise."

I would like people to use this thread to post other examples of boring advice. If you can, provide evidence and/or a plausible argument that your boring advice actually is useful, but I would prefer that you err on the side of boring but not necessarily useful in the name of more thoroughly searching a plausibly under-searched part of advicespace.

Upvotes on advice posted in this thread should be based on your estimate of the usefulness of the advice; in particular, please do not vote up advice just because it sounds deep or punchy.

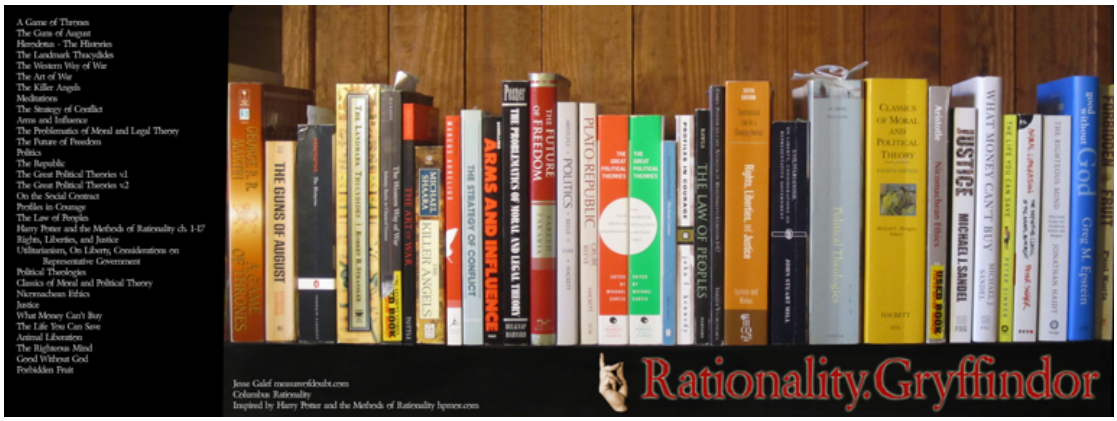
Harry Potter and the Methods of Rationality Bookshelves

A while back in the Columbus Rationality group, we started wondering: What books would the Harry Potter and the Methods of Rationality houses have in each of their libraries? We had fun categorizing different subjects:

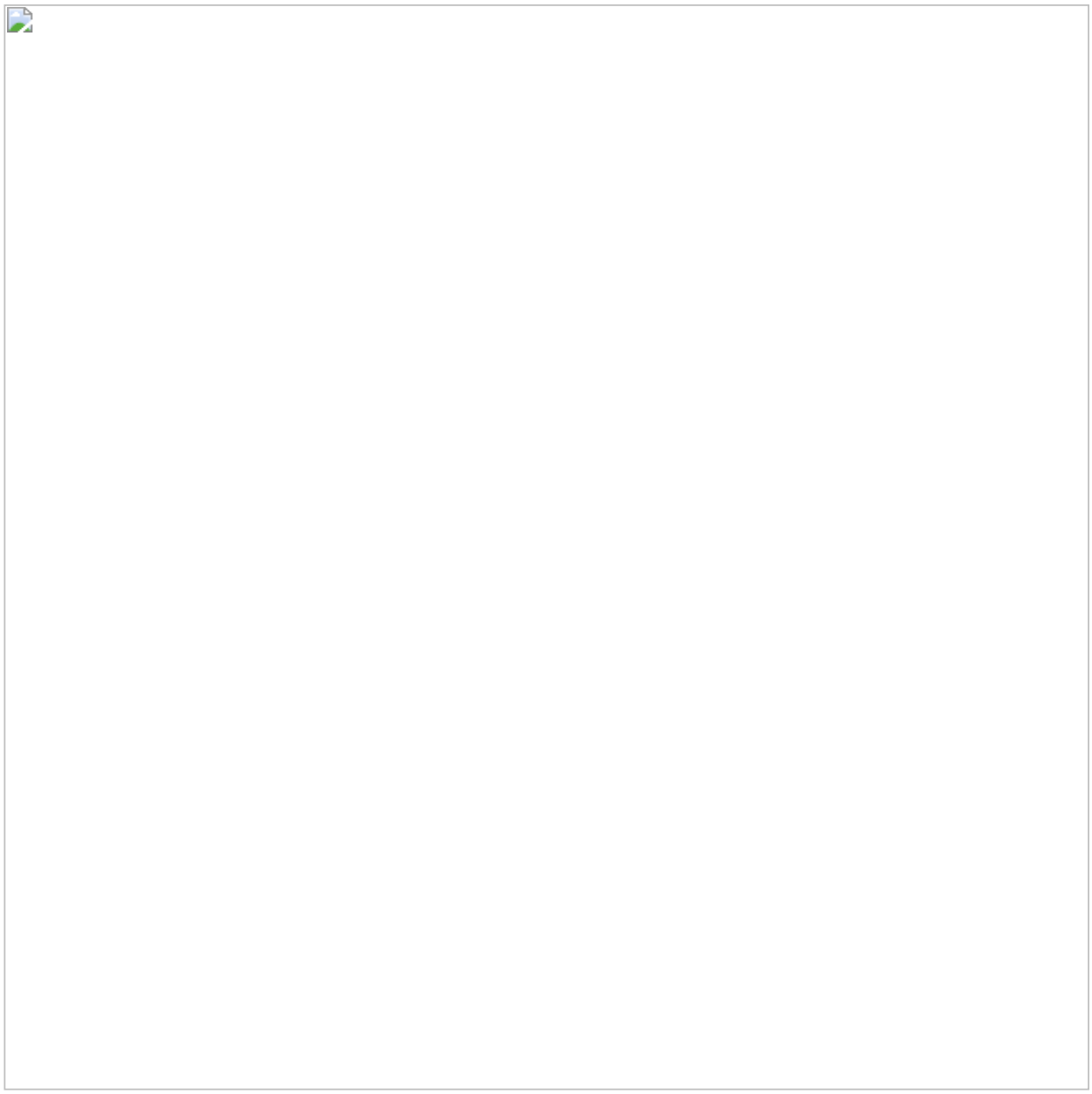
- Gryffindor - Combat, ethics, and justice
- Ravenclaw - Philosophy, cognitive science, and math
- Slytherin -Influence and power
- Hufflepuff - Happiness, productivity, and friendship

And so, I found myself taking all my books off their shelves this weekend and picking the best to represent each rationality!House and made them into Facebook cover-image-sized pictures. Click each image to see it larger, with a list on the left:

(first posted at [Measure of Doubt](measureofdoubt.com))




All Marketers Are Liars-Tell Stories
The Varnished Truth
Lying
Telling Lies
The Virtues of Our Vices
The Science of Fear
The Count of Monte Cristo
The International Thesaurus of Quotations
Emotional Intelligence
The 48 Laws of Power
59 Seconds
How to Argue
Harry Potter and the Methods of Rationality ch. 1-17
The Secret Life of Pronouns
Metaphors We Live By
Aristotle's Rhetoric
Yes! - 50 Scientifically Proven Ways to be Persuasive
Influence - The Psychology of Persuasion
The Screwtape Letters
How to Win Every Argument
Real Influence
How to Win Friends and Influence People
Thank You For Arguing
Made to Stick
Switch
Political Psychology
Micromotives and Macrobehavior
The Strategy of Conflict
Arms and Influence
The Republic
Rules for Radicals
Machiavelli's Selected Political Writings
Watchmen
Long for This World - The Strange Science of Immortality
WATCHMEN
MOORE
GIBBONS
LONG FOR THIS WORLD
JONATHAN WEINER
PLATO REPUBLIC
ARMS AND INFLUENCE
THE STRATEGY OF CONFLICT
Micromotives and Macrobehavior
Political Psychology
SWITCH
MADE TO STICK
THANK YOU FOR ARGUING
How to Win Friends & Influence People
REAL INFLUENCE
INFLUENCE the Psychology of Persuasion
RHETORIC
METAPHORS WE LIVE BY
HOW TO ARGUE
59 seconds
POWER
Emotional Intelligence
Daniel Goleman
COMPILED BY NICOLA THOMAS TRUPP
THESAURUS OF QUOTATIONS
The Science of Fear
PAUL EKMAN
Telling Lies
Rationality.Slytherin
Jesse Galef measureddoubt.com
Columbus Rationality
Inspired by Harry Potter and the Methods of Rationality hpmor.com



I’m always open to book recommendations and suggestions for good fits. What other books would be especially appropriate for each shelf?

Bayesian Adjustment Does Not Defeat Existential Risk Charity

(This is a long post. If you're going to read only part, please read sections 1 and 2, subsubsection 5.6.2, and the conclusion.)

1. Introduction

Suppose you want to give some money to charity: where can you get the most bang for your philanthropic buck? One way to make the decision is to use explicit expected value estimates. That is, you could get an unbiased (averaging to the true value) estimate of what each candidate for your donation would do with an additional dollar, and then pick the charity associated with the most promising estimate.

Holden Karnofsky of GiveWell, an organization that rates charities for cost-effectiveness, disagreed with this approach in two posts he made in 2011. This is a response to those posts, addressing the implications for existential risk efforts.

According to Karnofsky, high returns are rare, and even unbiased estimates don't take into account the reasons *why* they're rare. So in Karnofsky's view, our favorite charity shouldn't just be one associated with a high estimate, it should be one that supports the estimate with robust evidence derived from multiple independent lines of inquiry.[1] If a charity's returns are being estimated in a way that intuitively feels shaky, maybe that means the fact that high returns are rare should outweigh the fact that high returns were estimated, even if the people making the estimate were doing an excellent job of avoiding bias.

Karnofsky's first post, Why We Can't Take Expected Value Estimates Literally (Even When They're Unbiased), explains how one can mitigate this issue by supplementing an explicit estimate with what Karnofsky calls a "Bayesian Adjustment" (henceforth "BA"). This method treats estimates as merely noisy measures of true values. BA starts with a prior representing what cost-effectiveness values are out there in the general population of charities, then the prior is updated into a posterior in standard Bayesian fashion.

Karnofsky provides some example graphs, illustrating his preference for robustness. If the estimate error is small, the posterior lies close to the explicit estimate. But if the estimate error is large, the posterior lies close to the prior. In other words, if there simply aren't many high-return charities out there, a sharp estimate can be taken seriously, but a noisy estimate that says it has found a high-return charity must represent some sort of fluke.

Karnofsky does not advocate a policy of performing an *explicit* adjustment. Rather, he uses BA to emphasize that estimates are likely to be inadequate if they don't incorporate certain kinds of intuitions — in particular, a sense of whether all the components of an estimation procedure feel reliable. If intuitions say an estimate feels shaky and too good to be true, then maybe the estimate was noisy and the prior is more important. On the other hand, if intuitions say an estimate has taken everything into account, then maybe the estimate was sharp and outweighs the prior.

Karnofsky's second post, Maximizing Cost-Effectiveness Via Critical Inquiry, expands on these points. Where the first post looks at how BA is performed on a single charity at a time, the second post examines how BA affects the estimated relative values of different charities. In particular, it assumes that although the charities are all drawn from the same prior, they come with different estimates of cost-effectiveness. Higher estimates of cost-effectiveness come from estimation procedures with proportionally higher uncertainty.

It turns out that higher estimates aren't always more auspicious: an estimate may be "too good to be true," concentrating much of its evidential support on values that the prior already rules out for the most part. On the bright side, this effect can be mitigated via multiple independent observations, and such observations can provide enough evidence to solidify higher estimates despite their low prior probability.

Charities aiming to reduce existential risk have a potential claim to high expected returns, simply because of the size of the stakes. But if such charities are difficult to evaluate, and the prior probability of high expected values is low, then the implications of BA for this class of charities loom large.

This post will argue that competent efforts to reduce existential risk reduction are still likely to be optimal, despite BA. The argument will have three parts:

1. BA differs from fully Bayesian reasoning, so that BA risks double-counting priors.

2. The models in Karnofsky's posts, when applied to existential risk, boil down to our having prior knowledge that the claimed returns are virtually impossible. (Moreover, similar models without extreme priors don't lead to the same conclusions.)

3. We don't have such prior knowledge. Extreme priors would have implied false predictions in the past, imply unphysical predictions for the future, and are justified neither by our past experiences nor by any other considerations.

Claim 1 is not essential to the conclusion. While Claim 2 seems worth expanding on, it's Claim 3 that makes up the core of the controversy. Each of these concerns will be addressed in turn.

Before responding to the claims themselves, however, it's worth discussing a highly simplified model that will illustrate what Karnofsky's basic point is.

2. A Simple Discrete Distribution of Charitable Returns

Suppose you're considering a donation to the Center for Inventing Metawidgets (CIM), but you'd like to perform an analysis of the properties of metawidgets first.[2] Before the analysis, you're uncertain about three possibilities:

- With a probability of 4999 out of 10,000, metawidgets aren't even a thing. You can't invent what isn't a thing, so the return is 0.
- With a probability of 5000 out of 10,000, metawidgets are a thing with some reasonably good use, like repairing printers. The return in this case is 1.
- With a probability of 1 out of 10,000, metawidgets have extremely useful effects, like curing lung cancer. Then the return is 100.

If we now compute the expected value of a donation to CIM, it ends up as a sum of the following components:

- 0.4999 * 0 = **0** from the possibility that the return is 0
- 0.5 * 1 = **0.5** from the possibility that the return is 1
- 0.0001 * 100 = **0.01** from the possibility that the return is 100

In particular, the possibility of a modest return contributes 50 times the expected value of the possibility of an extreme return. The size of the potential return, in this case, didn't make up for its low probability.

But that's before you do an analysis that will give you some additional evidence about metawidgets. The analysis has the following properties:

- Whatever the true return is, 50% of the time the analysis is correct and gives you the correct answer.
- If the analysis is wrong, it picks one of the three possible answers uniformly at random.

What happens if the analysis says the return is 100?

To find the right probabilities to assign, we have to do Bayesian updating on this analysis result. The outcome of the analysis is four times as likely if the true value is 100 than if it is either 0 or 1. So the ratio of the expected value contributions changes from 50:1 to 50:4.

Applied to this case, Karnofsky's point is simply this: despite the analysis suggesting high returns, modest returns still come with higher expected value than high returns. High returns should be considered more probable after the analysis than before — we've observed a pretty good likelihood ratio of evidence in their favor — but high returns started out so improbable that even after receiving this bump, they still don't matter.

Now that we've seen the point in simplified form, let's begin a more detailed discussion.

3. The Role of BA

This section will add some critical notes on the concept of BA — notes that should apply whether the adjustment is performed explicitly or just used as a theoretical justification for listening to intuitions about the accuracy of particular estimates.

Before discussing the role of BA, let's guard against a possible misinterpretation. Karnofsky is not arguing against maximizing expected value. He is arguing against a particular estimation method he labels "Explicit Expected Value," which he considers to give inaccurate answers.

The Explicit Expected Value (EEV) method is simple: obtain an estimate of the true cost-effectiveness of an action, then act as if this estimate is the "true" cost-effectiveness. This "true" cost-effectiveness could be interpreted as an expected value itself.[3]

In contrast to EEV, Karnofsky advocates "Bayesian Adjustment." Bayesian reasoning involves multiplying a prior by a likelihood to find a posterior. In this case, the prior describes the charities that are out there in the population; the likelihood describes how likely different true values would have been to produce the given estimate; and the posterior represents our final beliefs about the charity's true cost-effectiveness. By looking at how common different effectiveness levels are, and how likely they would

have been to lead to the given estimate, we judge the probability of various effectiveness levels.

In the sense that we're updating on evidence according to Bayes' theorem, what's going on is indeed "Bayesian." But it's worth pointing out one difference between Karnofsky's adjustments and a fully Bayesian procedure: BA updates on a point estimate rather than on the full evidence that went into the point estimate.

This matters in two different ways.

First, the point estimate doesn't always carry all the available information. A procedure for generating a point estimate from a set of evidence could summarize different possible sets of evidence into the same point estimate, even though they favor different hypotheses. This sort of effect will probably be irrelevant in practice, but one might call BA "half-Bayesian" in light of it.

Second, and more importantly, there's a risk of misinterpreting the nature of the estimate. Karnofsky's model, again, assumes that estimates are "unbiased" — that conditional on any given number being the true value, if you make many estimates, they'll average out to that number. And if that's actually the case for the estimation procedure being used, then that's fine.

However, to the extent that an estimate took into account priors, that would make it "biased" toward the prior. As Oscar Cunningham comments:

> The people giving these estimates will have already used their own priors, and so you should only adjust their estimates to the extent to which your priors differ from theirs.

In the most straightforward case, the source simply gave his own Bayesian posterior mean. If you and the source had the same prior, then your posterior mean should be the source's posterior mean. After all, the source performed just the same computation that you would.

An old OvercomingBias post advises us to share likelihood ratios, not posterior beliefs. To be fair, in many cases communicating likelihood ratios for the whole space of hypotheses is impractical. One may instead want to communicate a number as a summary. (Even if one is making the estimate oneself, it may not be clear how one's brain came up with a particular number.) But it's important not to take a number that has prior information mixed in, and then interpret it as one that doesn't.

In less straightforward cases, maybe *part* of the prior was taken into account. For example, maybe your source shares your pessimism about the organizational efficiency of nonprofits, but not your pessimism in other areas. Even if your source informally ignored lines of reasoning that seemed to lead to an estimate that was "too good to be true," that is enough to make double-counting an issue.

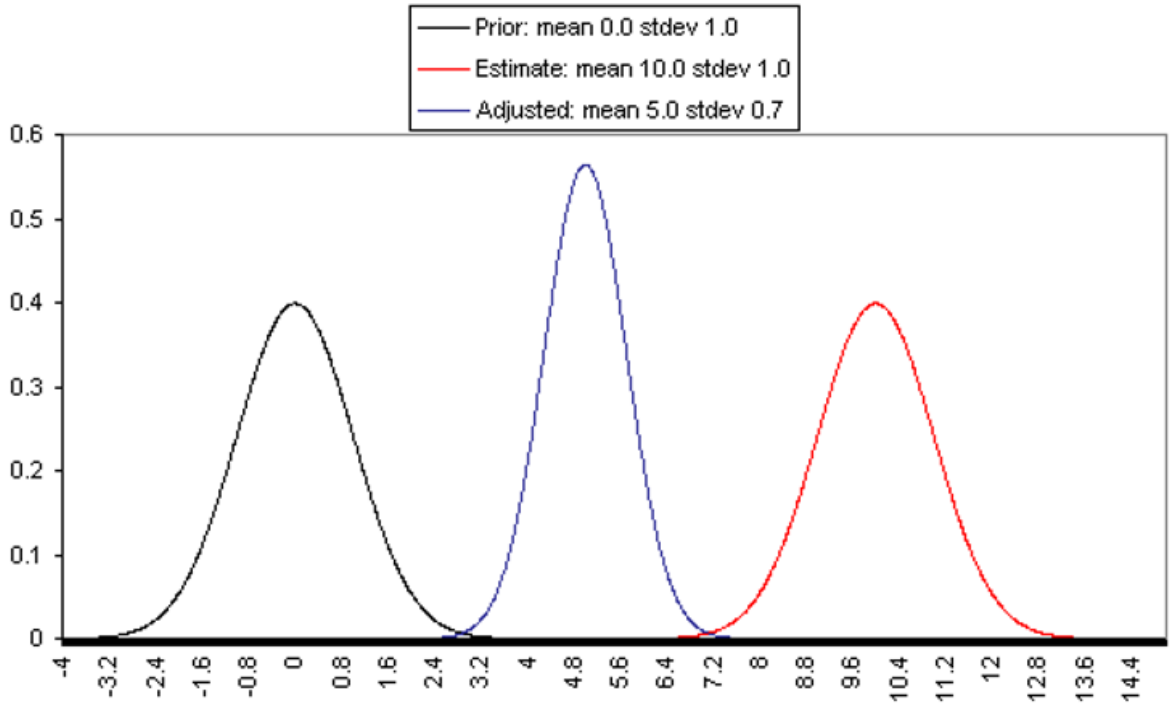
But to put this section in context, the appropriateness of BA isn't the most important disagreement with Karnofsky. Based on the considerations given here, performing an intuitive BA may well be better than going by an explicit estimate. Differences in priors have room to be far more important than just the results of (partially) double-counting them. So the more important part of the argument will be about which priors to use.

4. Probability Models

4.1: The first model

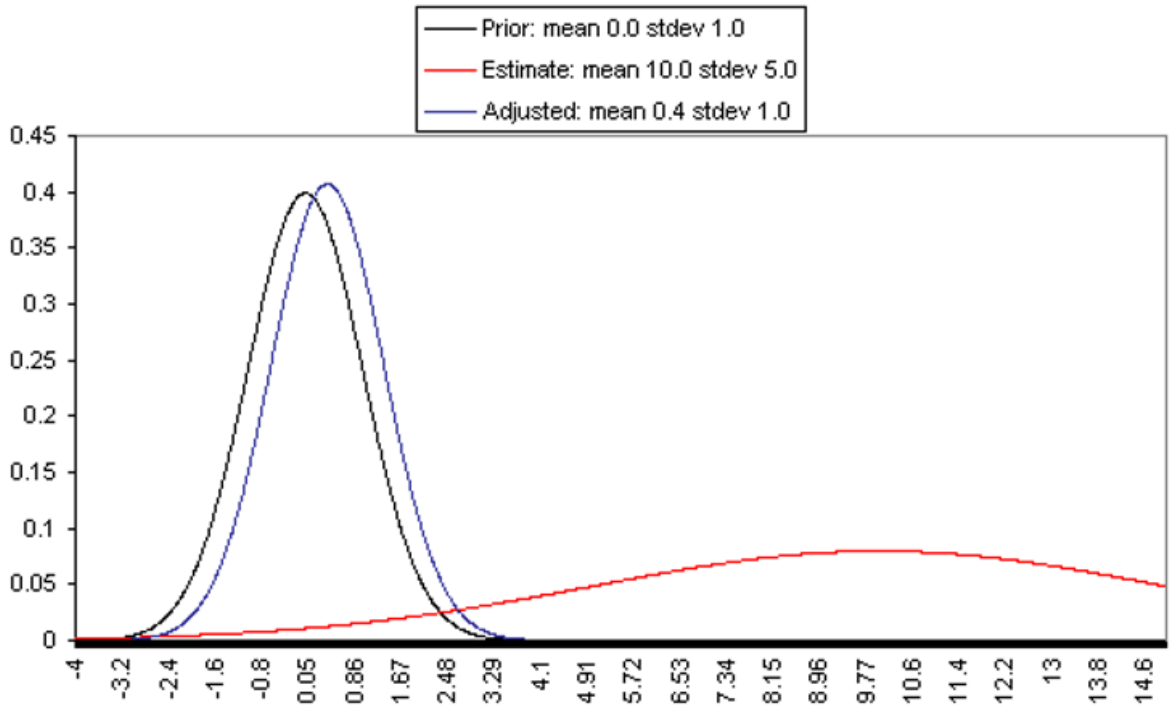
Karnofsky defends his conclusions with probabilistic models based on some mathematical calculations by Dario Amodei. This section will argue that these models only rule out optimal existential risk charity because the priors they assign to the relevant hypotheses are extremely low — in other words, because they virtually rule out extreme returns in advance.

In the model in Karnofsky's first post, it's easy to see the low priors. Consider the first example (the graphs are from Karnofsky's posts):



This example comes with some particular assumptions about parameters. The prior is normally distributed with mean 0 and standard deviation 1; the likelihood is normally distributed with mean 10 and standard deviation 1. As in the saying that "a Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule," the posterior ends up in the middle, hardly overlapping with either. As Eliezer Yudkowsky points out, this lack of overlap should practically *never happen*. When it does, such an event is a strong reason to doubt one's assumptions. It suggests that you should have assigned a different prior.

Or maybe, instead of the prior, it's the *likelihood* that you should have assigned differently — as one of the other graphs does:



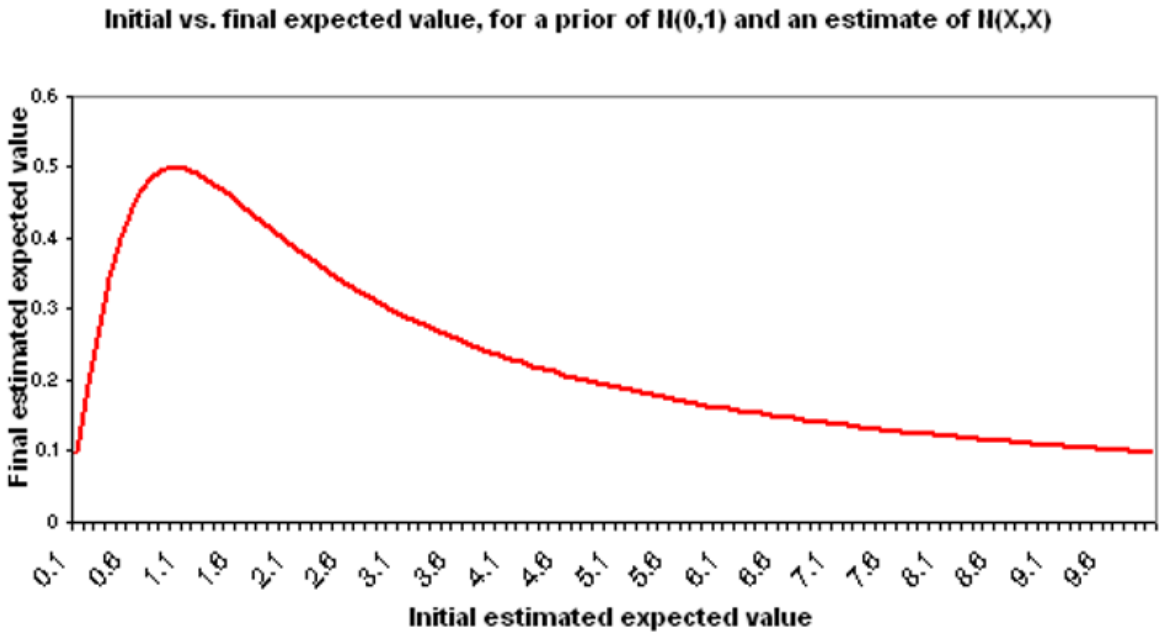
Here, the outcome makes some sense, because there's significant overlap. A high true cost-effectiveness would have been *more likely* to produce the estimate found, but a low true cost-effectiveness *could* have produced it instead. And the prior says the latter case, where the true cost-effectiveness is low, is far more likely — so the final best estimate, indeed, ends up not differing much from the initial best estimate.

Note, however, that this prior is extremely confident. The difference in probability density between the expected value and a value ten standard deviations out is a factor of e^{-50}, or about 10^{-22}. This number is so low it might as well be zero.

4.2: The second model

The second model builds on the first model, so many of the same considerations about extreme priors will carry over. This time, we're looking at a set of different estimates that we could be updating on like we did in the first model. For each of these, we take the expectation of the posterior distribution for the true cost-effectiveness, so we can put these expectations in a graph. After all, the expectation is the number that will factor into our decisions!

Here's one of the graphs, showing *initial estimates* on the x-axis and *final estimates* on the y-axis. The initial estimates are what we're performing a Bayesian update on, and the final estimates are the expectation value of the distribution of cost-effectiveness after updating:



So as initial estimates increase, the final estimate rises at first, but then slowly declines. High estimates are good up to a point, but when they become too extreme, we have to conclude they were a fluke.

As before, this model uses a standard normal prior, which means high true values have enormously smaller prior probabilities. Compared to this prior, the evidence provided by each estimate is minor. If the estimate falls one standard deviation out in the distribution, then it favors the estimate value over a value of zero by a likelihood ratio of the square root of e, or about 1.65. So it's no wonder that the tail end of high cost-effectiveness ends up irrelevant.

According to Karnofsky, this model illustrates that an estimate is safer to take at face value when evidence in its favor comes from multiple independent lines of inquiry. There are some calculations showing this — the more independent pieces of evidence for a given high value you gather, the more these together can overcome the "too good to be true" effect.

While multiple independent pieces of evidence are indeed better, it's important to emphasize that the relevant variable is simply the evidence's *strength*. Evidence can be strong because it comes from multiple directions, but it can also be strong because it just happens to be unlikely to occur under alternative hypotheses. If we have two independent observations that are both twice as likely to occur given cost-effectiveness 3 than cost-effectiveness 1, that's equally good as having a single observation that's four times as likely to occur given cost-effectiveness 3 than cost-effectiveness 1.

It's worth noting that if the multiple observations are all observations of one step in the process, and the other steps are left uncertain, there's a limit to how much multiple observations can make a difference.

4.3: Do the same calculations apply to log-normal priors?

Now that we've established that the models use low priors, can we evaluate whether the low priors are essential to the models' conclusions? Or are they just simplifying assumptions that make the math easier, but would be unnecessary in a full analysis?

One obvious step is to see if Karnofsky's conclusions hold up with log-normal models. Karnofsky states that the conclusions carry over qualitatively:

> the conceptual content of this post does not rely on the assumption that the value of donations (as measured in something like "lives saved" or "DALYs saved") is normally distributed. In particular, a log-normal distribution fits easily into the above framework

Assuming a log-normal prior, however, does change the mathematics. Graphs like those in Karnofsky's first post could certainly be interpreted as referring to the logarithm of cost-effectiveness, but the final number we're interested in is *the expected cost-effectiveness itself*. And if we interpret the graph as representing a logarithm, it's no longer the case that the point at the middle of the distribution gives us the expectation. Instead, values higher in the distribution matter more.

Guy Srinivasan points out that, for the same reason, log-normal priors would lead to different graphs in the second post, weakening the conclusion. To take the expectation of the logarithm and interpret that as the logarithm of the true cost-effectiveness is to bias the result downward.

If, instead of calculating e to the power of the expected value of the logarithm of cost-effectiveness, we calculate the expected value of cost-effectiveness directly, there's an additional term that increases with the standard deviation.

For an example of this, consider a normal distribution with mean 0 and standard deviation 1. If it represents the cost-effectiveness itself, we should take its expected value and find 0. But if it represents the logarithm of the cost-effectiveness, it won't do to take e to the power of the expected value, which would be 1. Rather, we add another ½ sigma (which in this case equals ½) before exponentiating. So the final expected cost-effectiveness ends up a factor sqrt(e) (= 1.65) larger — the most "average" value lies ½ to the right of the center of the graph.

While the mathematical point made here opposes Karnofsky's claims, it's hard to say how likely it is to be decisive in the context of the dilemmas that actually confront decision makers. So let's take a step back and directly face the question of how extreme these priors need to be.

4.4: Do priors need to be extreme?

As we've seen, Karnofsky's toy examples use extreme priors, and these priors would entail a substantial adjustment to EV estimates for existential risk charities. This adjustment would in turn be sufficient to alter existential risk charities from good ideas to bad ideas.[4]

The claim made in this section is: Karnofsky's models don't just *use* extreme priors, they *require* extreme priors if they are to have this altering effect. To determine whether this claim is true, one must check whether there are priors that aren't extreme, but still have the effect.[5]

And indeed, as pointed out by Karnofsky, there exist priors that (1) are far less extreme than the normal prior and (2) still justify a major adjustment to EV estimates for

existential risk charities. This is a sense in which his point qualitatively holds.

But the adjustment needs to be not just major, but large *enough* to turn existential risk charities from good ideas into bad ideas. This is difficult. Existential risk charities come with the potential for cost-effectiveness many orders of magnitude higher than that of the average charity. The normal prior succeeds at discounting this potential with its extreme skepticism, as may other priors. But if we can show that all the non-extreme priors justify an adjustment that may be large, but is not large enough to decide the issue, then that is a sense in which Karnofsky's point does not qualitatively hold.

And a prior can be far less extreme than the normal prior, while still being extreme. Do the log-normal prior and various even thicker-tailed priors qualify as "extreme," and do they entail sufficiently large adjustments? Rather than get hopelessly lost in that sort of analysis, let's just see what happens when one tries modeling real existential risk interventions as simple all-or-nothing bets: either they achieve some estimated reduction of risk, or the reasoning behind them fails completely.[6]

Suppose there's some estimate for the cost-effectiveness of a charity — call it E — and the true cost-effectiveness must be either 0 or E. You assign some probability p to the proposition that the estimate came from a true cost-effectiveness of E. This probability itself then comes from a prior probability that the estimate was E, and a likelihood ratio comparing at what rates true values of 0 and E create estimates of E.[7]

To find a ballpark number for what returns analyses are saying may be available from existential risk reduction (i.e., what value we should use for E), we can take a few different approaches.

One approach is to look at risks that are relatively tractable, such as asteroid impacts. It's estimated that impacts similar in size to that involved in the extinction of the dinosaurs occur about once every hundred million years. With the simplifying assumption that each such event causes human extinction, and that lesser asteroid events don't cause human extinction (or even end any existing lives), this translates to an extinction probability of one in a million for any given century. In other words, preventing all asteroid risk for a given century saves an expected 10^4 existing lives and an expected $1/10^6$ fraction of all future value.

A set of interventions funded in the past decade ruled out an imminent extinction-level impact at a cost of roughly $\$10^8$.[8]

According to this rough calculation, then, this program saved roughly one life plus a $1/(10^{10})$ fraction of the future for each $\$10^4$. Of course, future programs would probably be less effective.

For this to have been competitive with international aid ($\$10^3$ dollars per life saved), one only has to consider saving a 1 in 10^{10} fraction of humanity's entire future to be 10 times as important as saving an individual life. This is equivalent to considering saving humanity's entire future to be 10 times as important as saving all individual people living today. In a straightforward "astronomical waste" analysis, of course, it is far *more* important: enough so to compensate a high probability that the estimate is incorrect.

As an alternative to looking at tractable classes of risk for a cost-effectiveness estimate, we could look at the classes of existential risk that appear the most promising. AI risk, in particular, stands out. In a Singularity Summit talk, Anna Salamon estimated eight expected existing lives saved per dollar of AI risk research, or about

10^{-1} per existing life. Each existing life, again, also corresponds to a 10^{-10} fraction of our civilization's astronomical potential.

(There are a number of points where one could quibble with the reasoning that produced this estimate; cutting it down by a few orders of magnitude seems like it may not affect the underlying point too much. The main reason why there is an advantage here might be because we restricted ourselves to a limited class of charities for international aid, but not for existential risk reduction. In particular, the international aid charities we've used in the comparison are those that operate on an object level, e.g. by distributing mosquito nets, whereas the estimate in the talk refers to meta-level research about what object-level policies would be helpful.)

For such charities not to be competitive with international aid, just based on saving present-day lives alone, one would need to assign a probability that the estimate is correct of at most $1/10^4$. And as before, in a straightforward utilitarian analysis, the needed factor is much larger. This means that the probability that the estimate is correct could be far lower still.

Presumably the probability of an estimate of E given a true value of E is far greater than the probability of an estimate of E given a true value of 0. So the 10^4 or greater understates the extremeness of the priors you need. If your prior for existential risk-level returns is low because most charities are feel-good local charities, the likelihood ratio brings it back up a lot, because there aren't any feel-good local charities producing plausible calculations that say they're extremely effective.[9]

So one genuinely needs to find improbabilities that cut down the estimate by a large factor — although, depending on the specifics, one may need to bring in astronomical waste arguments to establish this point. Is it reasonable to adopt priors that have this effect?

5: Priors and their justification

5.1: Needed priors

To recapitulate, it turns out that if one uses the concepts in Karnofsky's posts to argue that (generally competent) existential risk charities are not highly cost-effective, this requires extreme priors. The least extreme priors that still create low enough posteriors are still fairly extreme.

Note that, for the argument to go through, it's not sufficient for the prior to be decreasing. A prior that doesn't decrease quickly enough doesn't even have a tail that's finite in size. Nor is it sufficient for the size of the prior's tail to be decreasing. It needs to at least decrease quickly enough to make up for the greater cost-effectiveness values we're multiplying by. For the expected value to even be finite a priori, with no evidence at all, the tail has to decrease more quickly than just at a minimum rate.

5.2: Possible justifications

Having argued that an attempt to defeat x-risk charities with BA requires a low prior — and that it therefore requires a justification for a low prior — let's look at possible approaches to such a justification.

One place to start looking could be in power laws. A lot of phenomena seem to follow power law distributions — although claims of power laws have also been criticized. The thickness of the tail depends on a parameter, but if, as this article) suggests, the parameter alpha tends to be near 1, then that gives one a specific thickness.

Another approach to justifying a low prior would be to say, "if such cost-effective strategies had been available, they would have been used up by now," like the proverbial $20 bill lying on the ground. (Here, it's a 20-util bill, which involves altruistic rather than egoistic incentives, but the point is still relevant.) Karnofsky has previously argued something similar.

For AI risk in particular, one might expect returns to have been driven down to the level of returns available for, e.g., asteroid impact prevention. If much higher returns are available for AI risk than other classes of risk, there must be some sort of explanation for why the low-hanging fruit there hasn't been picked.

Such an explanation requires us to think about the beliefs and motivations of those who fund measures to mitigate existential risks, although there may also simply be an element of random chance in which categories of threat get attention. Various differences between categories of risk are relevant. For example, AI risk is an area where relatively little expert consensus exists on how imminent the problem is, on what could be done to solve the problem, and even whether the problem exists. There are many reasons to believe that thinking about AI risk, compared to asteroids, is unusually difficult. AI risk involves thinking about many different academic fields, and offers many potential ways to become confused and end up mistaken about a number of complicated issues. Various biases could turn out to be a problem; in particular, the absurdity heuristic seems as though it could cause justified concerns to be dismissed early. Moreover, with AI risks, investment into global-scale risk is less likely to arise as a side effect of the prevention of smaller-scale disasters. Large asteroids pose similar issues to smaller asteroids, but human-level artificial general intelligence poses different issues than unintelligent viruses.

Of course, all these things are evidence against a problem existing. But they could also explain why, even in the presence of a problem, it wouldn't be acted upon.

5.3: Past experience as a justification for low priors

The main approach to justification of low priors cited by Karnofsky isn't any quantified argument, but is based on gut-level extrapolation from past experience:

> Even just a sense for the values of the small set of actions you've taken in your life, and observed the consequences of, gives you something to work with as far as an "outside view" and a starting probability distribution for the value of your actions; this distribution probably ought to have high variance, but when dealing with a rough estimate that has very high variance of its own, it may still be quite a meaningful prior.

It does not seem a straightforward task for a brain to extrapolate from its own life to global-scale efforts. The outcomes it has actually observed are likely to be a biased sample, involving cases where it can actually trace its causal contribution to a relatively small event. In particular, of course, a brain hasn't had any opportunity to observe effects persisting for longer than a human lifetime.

Extrapolating from the mundane events your brain has directly experienced to far out in the tail, where the selection of events has been highly optimized for utilitarian

impact, is likely to be difficult.

“Black swan” type considerations are relevant here: if you’ve seen a million white swans in a row in the northern hemisphere, that might entitle you to assign a low probability that the first swan you see in the southern hemisphere will be non-white, but it doesn’t entitle you to assign a one-in-a-million probability. In just the same way, if you’ve seen a million inefficient charities in a row when looking mostly at animal charities, that doesn’t entitle you to assign a one-in-a-million probability to a charity in the class of international aid being efficient. Maybe things will just be fundamentally different.

But it can be argued that we have already had some actual observations of existential risk-scale interventions. And indeed, Karnofsky says elsewhere that past claims of enormous cost-effectiveness have failed to pan out:

> I think that speaking generally/historically/intuitively, the number of actions that a back-of-the-envelope calc could/would have predicted enormous value for is high, and the number that panned out is low. So a claim of enormous value is sufficient to make me skeptical. In other words, my prior isn’t so wide as to have little regressive impact on claims like "1% chance of saving the world."

One can argue the numbers: exactly how many actions seemed enormously valuable in the way AI risk reduction seems to? Exactly how few of them panned out? Some examples one might include in this category are religious claims about the afterlife or the end times, particularly leveraged ways of creating permanent social change, or ways to intervene at important points in nuclear arms races. But in general, if your high estimate of cost-effectiveness for an organization is based on, say, a 10% chance that it would visibly succeed at achieving enormous returns over its lifetime, then just a few such failures provide only moderate evidence against the accuracy of the estimate. And as we’ve seen, for the regressive impact created by Karnofsky’s priors to make a difference, it needs to be not just substantial, but enormous.

5.4: Intuitions suggesting extremely low priors are unreasonable

To get a feel for how extreme some of these priors are, consider what they would have predicted in the past. As Carl Shulman says:

> [I]t appears that one can save lives hundreds of times more cheaply through vaccinations in the developing world than through typical charity expenditures aimed at saving lives in rich countries, according to experiments, government statistics, etc.
>
> But a normal distribution (assigns) a probability of one in tens of thousands that a sample will be more than 4 standard deviations above the median, and one in hundreds of billions that a charity will be more than 7 standard deviations from the median.

In other words, with a normal prior, the model assigns extremely small probabilities to events that have, in fact, happened. With a log-normal prior, the problem is not as bad. But as Shulman points out, such a prior still makes predictions for the future that are difficult to square with physics — difficult to square with the observation that existential disasters seem possible, and at least some of them are partly mediated by technology. As a reductio ad absurdum of normal and log-normal priors, he offers a “charity doomsday argument”:

If we believed a normal prior then we could reason as follows:

1. If humanity has a reasonable chance of surviving to build a lasting advanced civilization, then some charity interventions are immensely cost-effective, e.g. the historically successful efforts in asteroid tracking.

2. By the normal (or log-normal) prior on charity cost-effectiveness, no charity can be immensely cost-effective (with overwhelming probability).

Therefore,

1. Humanity is doomed to premature extinction, stagnation, or an otherwise cramped future.

In Karnofsky's reactions to arguments such as these, he has emphasized that, while his model may not be realistic, there is no better model available that leads to different conclusions:

> You and others have pointed out that there are ways in which my model doesn't seem to match reality. There are definitely ways in which this is true, but I don't think pointing this out is - in itself - much of an objection. All models are simplifications. They all break down in some cases. The question is whether there is a better model that leads to different big-picture implications; no one has yet proposed such a model, and my intuition is that there is not one.

But the flaw identified here — that the prior in Karnofsky's models cannot be convinced of astronomical waste — isn't just an accidental feature of simplifying reality in a particular way. It's a flaw present in any scheme that discounts the implications of astronomical waste through priors. Whatever the probability for the existence of preventable astronomical waste is, in expected utility calculations, it gets multiplied by such a large number that unless it starts out extremely low, there's a problem.

As a last thought experiment suggesting the necessary probabilities are extreme, suppose that in addition to the available evidence, you had a magical coin that always flipped heads if astronomical waste were real and preventable — but that was otherwise fair. If the coin came up heads dozens of times, wouldn't you start to change your mind? If so, unless your intuitions about coins are heavily broken, your prior must not in fact be so extremely small as to cancel out the returns.

5.5: Indirect effects of international aid

There is a possible way to argue for international aid over existential risk reduction based on priors without requiring a prior so small as to unreasonably deny astronomical waste. Namely, one could note that international aid itself has effects on astronomical waste. Then international aid is on a more equal level with existential risk, no matter how large the numbers for astronomical waste turn out to be.

Perhaps international aid has effects hastening the start of space colonization. Earlier space colonization would prevent whatever astronomical waste takes place during the interval between the point where space colonization actually happens, and the point where it would otherwise have happened. This could conceivably outweigh the astronomical waste from existential risks even if such risks aren't astronomically improbable.

Do we have a way to evaluate such indirect effects on growth? The argument goes as follows: international aid saves people's lives, saving people's lives increases economic growth, economic growth increases the speed of development of the required technologies, and this decreases the amount of astronomical waste. However, as Bostrom points out in his paper on astronomical waste, safety is still a lot more important than speed:

> If what we are concerned with is (something like) maximizing the expected number of worthwhile lives that we will create, then in addition to the opportunity cost of delayed colonization, we have to take into account the risk of failure to colonize at all. … Because the lifespan of galaxies is measured in billions of years, whereas the time-scale of any delays that we could realistically affect would rather be measured in years or decades, the consideration of risk trumps the consideration of opportunity cost. For example, a single percentage point of reduction of existential risks would be worth (from a utilitarian expected utility point-of-view) a delay of over 10 million years.

A more recent analysis by Stuart Armstrong and Anders Sandberg emphasizes the effect of galaxies escaping over the cosmic event horizon: the more we delay colonization, and the more slowly colonization happens, the more galaxies go permanently out of reach. Their model implies that we lose about a galaxy per year of delaying colonization at light speed, or about a galaxy every fifty years of delaying colonization at half light speed. This is out of, respectively, 6.3 billion and 120 million total galaxies reached.

So a year's delay wastes only about the same amount of value as a one-in-several-billion chance of human extinction. That means safety is usually more important than delay. For delay to outweigh safety requires a highly confident belief in the proposition that we can affect delay but not safety.

Does this give us a way to estimate the indirect returns of saving one person's life in the Third World?

Since it's probably good enough to estimate to within a few orders of magnitude, we'll make some very loose assumptions.

Suppose a Third World country with a population of 100 million makes a total difference of one month in the timing of humanity's future colonization of space. Then a single person in that country makes an expected difference of 1/(1200 million) years — equivalent to a one-in-billions-of-billions chance of human extinction.

If saving the person's life is the result of an investment of $\$10^3$, then to claim the astronomical waste returns are similar to those from preventing existential risk, one must claim an existential risk intervention of $\$10^6$ would have a chance of one in millions of billions of preventing an existential disaster, and an intervention of $\$10^9$ would have a chance of one in thousands of billions.

There are some caveats to be made on both sides of the argument. For example, we assumed that preventing human extinction has billions of times the payoff of delaying space colonization for a year; but what if the bottleneck is some other resource than what's being wasted? In that case, it could be that, if we survive, we can get a lot more value than billions of times what is lost through a year's waste. And if one (naively?) took the expectation value of this "billions" figure, one would probably end up with something infinite, because we don't know for sure what's possible in physics.

Increased economic growth could have effects not just on timing, but on safety itself. For example, economic growth could increase existential risk by speeding up dangerous technologies more quickly than society can handle them safely, or it could decrease existential risk by promoting some sort of stability. It could also have various small but permanent effects on the future.

Still, it would seem to be a fairly major coincidence if the policy of saving people's lives in the Third World were also the policy that maximized safety. One would at least expect to see more effect from interventions targeted specifically at speeding up economic growth. An approach to foreign aid aimed at maximizing growth effects rather than near-term lives or DALYs saved would probably look quite different. Even then, it's hard to see how economic growth could be the policy that maximized safety unless our model of what causes safety were so broken as to be useless.

Throughout this analysis, we've been assuming a standard utilitarian view, where the loss of astronomical numbers of future life-years is more important than the deaths of current people by a correspondingly astronomic factor. What if, at the other extreme, one only cared about saving as many people as possible from the present generation? Then delay might be more important: in any given year, a nontrivial fraction of the world population dies. One could imagine a speedup of certain technologies causing these technologies to save the lives of whoever would have died during that time.

Again, we can do a very rough calculation. Every second, 1.8 people die. So if, as above, saving a life through malaria nets makes a difference in colonization timing of 1/(1200 million) years or 25 milliseconds, and if hastening colonization by one second saves those 1.8 lives, the additional lives saved through the speedup are only 1/40 of the lives saved directly by the malaria net.

Since we're dealing with order-of-magnitude differences, for this 1/40 to matter, we'd need to have underestimated it by orders of magnitude. What we'd have to prove isn't just that lives saved through speedup outnumber lives saved directly; what we'd have to prove is that lives saved through speedup outnumber lives saved through alternative uses of money. As we saw before, on top of the 1/40, there are still another four orders of magnitude or so between estimates of the returns in current lives saved through AI risk reduction and international aid.

One may question whether this argument constitutes a "true rejection" of the cost-effectiveness of existential risk reduction: were international aid charities really chosen *because* they increase economic growth and thereby speed up space colonization? If one were optimizing for that criterion, presumably there would be more efficient charities available, and it might be interesting to look at whether one could make a case that they save more current people than AI risk reduction. One would also need to have a reason to disregard astronomical waste.

5.6: Pascal's Mugging and the big picture

Let's take a more detailed look at the question of whether reasonable priors, in fact, bring the expected returns of the best existential risk charities down by a sufficient factor. Karnofsky states a general argument:

> But as stated above, I believe even most power-law distributions would lead to the same big-picture conclusions. I believe the crucial question to be whether the prior probability of having impact >=X falls faster than X rises. My intuition is that for any reasonable prior distribution for which this is true, the big-picture conclusions

of the model will hold; for any prior distribution for which it isn't true, there will be major strange and problematic implications.

In defending the idea that existential risk reduction has a high enough probability of success to be a good investment, we have two options:

1. Use a prior with a tail that decreases faster than 1/X, and argue that the posterior ends up high enough anyway.

2. Use a prior with a tail that decreases slower than 1/X, and argue that there are no strange implications; or that there are strange implications but they're not problematic.

Let's briefly examine both of these possibilities. We can't do the problem full numerical justice, but we can at least take an initial stab at answering the question of what alternative models could look like.

5.6.1: Rapidly shrinking tails

First, let's look at an example where the prior probability of impact at least X falls *faster* than X rises. Suppose we quantify X in terms of the number of lives that can be saved for one million dollars. Consider a Pareto distribution (that is, a power law) for X, with a minimum possible value of 10, and with alpha equal to 1.5 so that the density for X decreases as $X^{-5/2}$, and the probability mass of the tail beyond X decreases as $X^{-3/2}$. Now suppose international aid claims an X of at least 1000 and existential risk reduction claims an X of at least 100,000. Then there's a 1 in 1000 prior for the international aid tail and a 1 in 1000000 prior for the existential risk tail.

A one in a million prior sounds scary. However:

- Those million charities would consist almost entirely of obviously non-optimal charities. Just knowing the general category of what they're trying to do would be enough to see they lacked extremely high returns. Picking the ones that are even mildly reasonable candidates already involves a great deal of optimization power.

- You wouldn't need to identify the one charity that had extremely good returns. For purposes of getting a better expected value, it would be more than sufficient to narrow it down to a list of one hundred.

- Presumably, some international aid charities manage to overcome that 1 in 1000 prior, and reach a large probability. If reasoning can pick out the best charity in a thousand with reasonable confidence, then maybe once those charities are picked out, reasoning can take a useful guess at which one is the best in a thousand of *these* charities.

- Overconfidence studies have trained us to be wary of claims that involve 99.99% certainty. But we should be wary of a confident prior just as we should be wary of a confident likelihood. It's easy to make errors when caution is applied in only one direction. As a further "intuition pump," suppose you're in a foreign country and you meet someone you know. The prior odds against it being that person may be billions to one. But when you meet them, you'll soon have strong enough evidence to attain nearly 100% confidence — despite the fact that this takes a likelihood ratio of billions.

So in sum, it seems as though even with a prior that declines fairly quickly, an analysis could still reasonably judge existential risk-level returns to be the most important. A quickly declining prior can still be overcome by evidence — and the amount of evidence needed drops to zero as the size of the tail gets closer to decreasing at a speed of 1/X. Again, just because an effect exists in a qualitative sense, that doesn't mean that, in practice, it will affect the conclusion.

5.6.2: Slowly shrinking tails

Second, let's consider prior distributions where the probability of impact at least X falls slower than X rises. One example of where this happens is a power law with an alpha lower than 1. But priors implied by Solomonoff induction also behave like this. For example, the probability they assign to a value of 3^^^3 is much larger than 1/(3^^^3), because the number can be produced by a relatively short program. Most values that large have negligibly small probabilities, because there's no short program for them. But some values that large have higher probabilities, and end up dominating any plausible expected value calculation starting from such a prior. [10]

This problem is known as "Pascal's Mugging," and has been discussed extensively on LessWrong. Karnofsky considers it a reason to reject any prior that doesn't decrease fast enough. But there are a number of possible ways out of the problem, and not all of them change the prior:

- Adopting a bounded utility function (with the right bound and functional form) can make it impossible for the mugger to make promises large enough to overcome their improbability.
- One could bite the bullet by accepting that one should pay the mugger — or rather that more plausible "muggers," in the form of infinite physics, say, may come along later.
- If the positive and negative effects of giving in to muggers are symmetrical on expectation, then they cancel out... but why would they be symmetrical?
- Discounting the utility of an effect by the algorithmic complexity of locating it in the world implies a special case of a bounded utility function.
- One could ignore the mugger for game-theoretical reasons... however, the hypothetical can be modified to make game theory irrelevant.
- One could justify a quickly declining prior using anthropic reasoning, as in Robin Hanson's comment: statistically, most agents can't determine the course of a vast number of agents' lives. However, while this is a plausible claim about anthropic reasoning, if one has uncertainty about what is the right account of anthropic reasoning, and if one treats this uncertainty as a regular probability, then the Pascal's Mugging problem reappears.
- One could justify a quickly declining prior some other way.

With regard to the last option, one does need some sort of justification. A probability doesn't seem like something you can choose based on whether it implies reasonable-sounding decisions; it seems like something that has to come from a model of the world. And to return to the magical coin example, would it really take roughly log(3^^^3) heads outcomes in a row (assuming away things like fake memories) to convince you the mugger was speaking the truth?

It's worth taking particular note of the second-to-last option, where a prior is justified using anthropic reasoning. Such a prior would have to be quickly declining. Let's explore this possibility a little further.

Suppose, roughly speaking, that before you know anything about where you find yourself in the universe, you expect on average to decisively affect one person's life. Then your prior for your impact should have an expectation value less than infinity — as is the case for power laws with alpha greater than 1, but not alpha smaller than 1. Of course, the number of lives a rational philanthropist affects is likely to be larger than the number of lives an average person affects. But if some people are optimal philanthropists, that still puts an upper bound on the expectation value. Likewise, if most things that could carry value aren't decision makers, that's a reason to expect greater returns per decision maker. Still, it seems like there would be some constant upper bound that doesn't scale with the size of the universe.

In a world where whoever happens to be on the stage at a critical time gets to determine its long-term contents, there's a large prior probability that you're causally downstream of the most important events, and an extremely small prior probability that you live exactly at the critical point. Then suppose you find yourself on Earth in 2013, with an apparent astronomical-scale future still ahead, depending on what happens between now and the development of the relevant technology. This seems like it should cause a strong update from the anthropic prior. It's possible to find ways in which astronomical waste could be illusory, but to find them we need to look in odd places.

- One candidate hypothesis is the idea that we're living in an ancestor simulation. This would imply astronomical waste was illusory: after all, if a substantial fraction of astronomical resources were dedicated toward such simulations, each of them would be able to determine only a small part of what happened to the resources. This would limit returns. It would be interesting to see more analysis of optimal philanthropy given that we're in a simulation, but it doesn't seem as if one would want to predicate one's case on that hypothesis.

- Other candidate hypotheses might revolve around interstellar colonization being impossible even in the long run for reasons we don't currently understand, or around the extinction of human civilization becoming almost inevitable given the availability of some future technology.

- As a last resort, we could hypothesize nonspecific insanity on our part, in a sort of majoritarian hypothesis. But it seems like assuming that we're insane and that we have no idea *how* we are insane undermines a lot of the other assumptions we're using in this analysis.

If Karnofsky or others would propose other such factors that might create the illusion of astronomical waste, or if they would defend any of the ones named, spelling them out and putting some sort of rough estimate or bounds on how much they tell us to discount astronomical waste seems like it would be an important next move in the debate.

It may be a useful reframing to see things from a perspective like Updateless Decision Theory. The question is whether one can get more value from controlling structures that — in an astronomical-sized universe — are likely to exist many times, than from an extremely small probability of controlling the whole thing.

6. Conclusion

BA doesn’t justify a belief that existential risk charities, despite high back-of-envelope cost-effectiveness estimates, offer low or mediocre expected returns.

We can assert this without having to endorse claims to the effect that one must support (without further research) the first charity that names a sufficiently large number. There are other considerations that defeat such claims.

For one thing, there are multiple charities in the general existential risk space and potentially multiple ways of donating to them; even if there weren’t, more could be created in the future. That means we need to investigate the effectiveness of each one.

For another thing, even if there were only one charity with great potential returns in the area, you’d have to check that marginal money wasn’t being negatively useful, as Karnofsky has argued is indeed the case for MIRI (because the "Friendly AI" approach is unnecessarily dangerous, according to Karnofsky).

Systematic upward bias, not just random error, is of course likely to play a role in organizations’ estimates of their own effectiveness.

And finally, some other consideration, not covered in these posts, could prove either that existential risk reduction doesn’t have a particularly high expected value, or that we shouldn’t maximize expected value at all. (Bounded utility functions are a special case of not maximizing expected value, if “value” is measured in e.g. DALYs rather than utils.) Note, however, that Karnofsky himself has not endorsed the use of non-additive metrics of charitable impact.

MIRI, in choosing a strategy, is not gambling on a tiny probability that its actions will turn out relevant. It’s trying to affect a large-scale event — the variable of whether or not the intelligence explosion turns out safe — that will eventually be resolved into a “yes” or “no” outcome. That every individual dollar or hour spent will fail to have much of an effect by itself is an issue inherent to pushing on large-scale events. Other cases where this applies, and where it would not be seen as problematic, are political campaigns and medical research, if the good the research does comes from a few discoveries spread among many labs and experiments.

The improbability here isn’t in itself pathological, or a stretch of expected value maximization. It might be pathological if the argument relied on further highly improbable “just in case” assumptions, for example if we were almost certain that AI is impossible to create, or if we were almost certain that safety will be ensured by default. But even though “if there’s even a chance” arguments have sometimes been made, MIRI does not actually believe that there’s an additional factor on top of that inherent per-dollar improbability that would make it so that all its efforts are probably irrelevant. If it believed that, then it would pick a different strategy.

All things considered, our evidence about the distribution of charities is compatible with AI being associated with major existential risks, and compatible with there being low-hanging fruit to be picked in mitigating such risks. Investing in reducing existential risk, then, can be optimal without falling to BA — and without strange implications.

Notes

This post was written by Steven Kaas and funded by MIRI. My thanks for helpful feedback from Holden Karnofsky, Carl Shulman, Nick Beckstead, Luke Muehlhauser, Steve Rayhawk, and Benjamin Noble.

[1] It's worth noting, however, that Karnofsky's vision for GiveWell is to provide donors with the best giving opportunities that can be found, not necessarily the giving opportunities whose ROI estimates have the strongest evidential backing. So, for Karnofsky, strong evidential backing is a means to the end of finding the best interventions, not an end in itself. In Givewell's January 24th, 2013 board meeting (starting at 24:30 in the MP3 recording), Karnofsky said:

"The way ["GiveWell 2", a possible future GiveWell focused on giving opportunities for which strong evidence is less available than is the case with GiveWell's current charity recommendations] would prioritize [giving] opportunities would involve... a heavy dose of personal judgment, and a heavy dose of... "Well, we have laid out our reasons of thinking this. Not all the reasons are things we can prove, but... here's the evidence we have, here's what we do know, and given the limited available information here's what we would guess." We actually do a fair amount of that already with GiveWell, but it would definitely be more noticeable and more prominent and more extreme [in GiveWell 2]...

...What would still be "GiveWell" about ["GiveWell 2"] is that I don't believe that there's another organization that's out there that is publicly writing about what it thinks are the best giving opportunities and why, and... comparing all the possible things you might give to... It's basically a topic of discussion that I don't believe exists right now, and... we started GiveWell to start that discussion in an open, public way, and we started in a certain place, but *that* and not evidence... has always been the driving philosophy of GiveWell, and our mission statement talks about expanding giving opportunities, it doesn't talk about evidence."

[2] Technically, the prior is usually not about a specific charity that we already have information about, but about charities in general. I give an example of a specific fictional charity because I figured that would be more clarifying, and the math works as long as you're using an estimate to move from a state of less information to a state of more information.

[3] At least in the sense that it might still average over, say, quantum branching and chaotic dynamics. But the "true value" would at least be based on a full understanding of the problem and its solutions.

[4] Of course, it may be the case that particular charities working on existential risk reduction fail to pursue activities that *actually* reduce existential risk — that question is separate from the questions we have the space to examine here.

[5] For this section, by "extreme priors" I just mean something like "many zeroes." Does the prior say that what some of us think of as always having been a live hypothesis actually started out as hugely improbable? Then it's "extreme" for my purposes. Once it's been established that only extreme priors let the point carry through, one can then discuss whether a prior that's "extreme" in this sense may nonetheless be justified. This is what the next section will be devoted to. The separation between these two points forces me to use this rather artificial concept of "extreme," where an analysis would ideally just consider what priors are reasonable and how Karnofsky's point works with them. Nonetheless, I hope it makes things clearer.

[6] It would be nice to have some better examples of the overall point, but these were the examples that seemed maximally illustrative, clear, and concise given time and space constraints.

[7] This estimate, technically, isn't unbiased. If the true value is E, the estimate will average lower than E, and if the true value is 0, the estimate will average higher than 0. But this shouldn't matter for the illustration.

[8] To be sure, if an asteroid had been on its way, we would have also needed to pay the cost of deflecting it. But this possibility was extremely improbable. As long as the cost of deflection wouldn't have been much more than 10^{14}, this doesn't increase the expected cost by orders of magnitude.

[9] There are some points to be made here about causal screening, and also that it's unnatural to think of the prior as being on effectiveness, rather than on things that cause both effectiveness and low priors, unless effectiveness is a thing that causes low priors, for example because people have picked up all the low-hanging fruit off the ground. But due to time and space concerns, I have left those points out of this document.

[10] A more complete argument would involve looking at how often a given structure would be repeated with what probability in a simplicity-weighted set of universes, but the general point is the same.

Schelling Day: A Rationalist Holiday

The Big Idea

—Holidays are awesome.
—Getting to know people is awesome.
—Schelling Day is a holiday about getting to know people.
—The Boston group will celebrate Schelling day.
—Hopefully other cities will too.

Why Are We Doing This?

Getting to know people—really, truly getting to know people—is hard. You have to spend a huge amount of time with them, of course, but that's the easy part. Spending time with people is fun! The challenging part is opening yourself up. Sharing your fondest hopes and deepest fears is a powerful way to make connections, but exposing your soul like that terrifying. Worse, it's awkward. There's no socially appropriate time to bring up stuff like that. I've talked to a bunch of people who wish there were more opportunities for that sort of sharing, but initiating it is risky. Even when everything works out beautifully, getting it started feels stressful and not-fun.

What if we could set aside a time where sharing like that is not merely accepted, but expected? Historically, this doesn't seem too hard to do. As soon as people are in a context where everyone agrees that sharing is normal (e.g. an Alcoholics Anonymous meeting, or a conversation with a therapist), the stigma and self-consciousness don't hold people back nearly as much.

All we need is an arbitrary time when we agree to change the social rules, and we're set! In other words, we need a Schelling point. April 14th, the birthday of Thomas Schelling, is as good a time as any.

I'm creating a ritual around this, for two reasons. First, an established structure makes it easier to do something that feels difficult or strange. Second, in my experience, adding ritual to powerful, true statements makes them even more powerful.

When Are We Doing This?

Schelling Day is April 14th, which is a Sunday this year. The Boston group will be holding our celebration at 2:30. The ritual will begin sharply at 2:45, so please be on time.

Please RSVP to the meetup if you're coming. We're allowed to have up to 20 people in the space, and we'll be using the Meetup site to track this. It's fine if you RSVP at 2:00 on the day of the event, so long as you don't put us above the limit. If for some reason

you don’t want your RSVP to be public, PM me and I’ll reserve you a spot anonymously.

There will be a potluck dinner. Everyone who brings a dish gets two Rationality Points.

What Are We Doing?

Everyone sits in a semicircle. At the focal point are two tables. On the first table are five small bowls of delicious snacks. Eating the delicious snacks at this stage is VERBOTEN. On the second table is a single large, empty bowl.

Everyone will have a six-sided die.

Everyone will have a chance to speak, or to not speak. When it’s your turn, roll your die. Showing the result to others is VERBOTEN.

If your die shows a six, you MUST speak. If your die shows a one, you MAY NOT speak. Otherwise, you choose whether or not to speak. The die is to provide plausible deniability. Attempting to guess whether someone’s decision was forced by the die roll is VERBOTEN.

If you speak, take 1-5 minutes* to tell the group about one of your secret Joys, Struggles, Hopes, Confessions, or Something Else Important, as described below. Then scoop some food from the appropriate bowl and put it into the larger bowl.

Struggles (Chocolate):
Flaws, interpersonal drama, professional challenges, stuff you’d say to a therapist

Joys (Raspberries):
Passions, guilty pleasures, “I love you guys” speeches

Confessions (Pretzels):
Burdens, personal secrets, things you’re tired of hiding, stuff you’d say to a priest

Hopes (Raisins):
Goals, wishes, deepest desires, crazy schemes

Something Else (Trail mix):
Because trying to make an exhaustive list would be silly.

After your speak, or after you choose not to speak, the person to your left rolls their die and the process repeats.

Once everyone has had a chance to speak or not, take five minutes* to stretch, then do the same thing again.

After that, take five minutes to stretch, then begin the BONUS ROUND.

The BONUS ROUND is like the first two rounds, with one exception. If you haven't spoken yet, do not roll your die. You MUST speak.

Then What?

We'll pass around the bowl of snacks we've assembled from our accumulated revelations until everything is eaten.

Depending on the timing, the emotional state, and our patience, we might or might not have another round or two.

After that, dinner! The rest of the time will be for eating and socializing. We'll break into smaller groups and follow up on the things we said during the ritual. Asking questions about what someone said is actively encouraged! (There is no obligation to answer. "I would prefer not to talk about that" is a completely acceptable response.) Err on the side of asking an awkward question; if you're over the line, the other person will simply decline to answer, and no harm done. Judging people, or explaining why their revelations were wrong, is of course VERBOTEN—unless someone specifically asks for feedback, in which case be honest but don't be a jerk. We'll get the potluck dishes people brought, and we'll eat, drink, and be merry.

*I'll be using a timer! I don't want to be a jerk, but I want to keep things moving.

UPDATE: My review of the event is here.

Rationality Habits I Learned at the CFAR Workshop

Recently Leah Libresco asked attendees at the January CFAR Workshop, "What habits have people installed after workshops?" and that got me thinking that now was a good time to write up and review what I learned (or learned and already forgot). I thought that might be of some interest to folks here, and this is what follows.

What I Learned and Implemented

The most immediately useful thing I learned was the *Pomodoro Technique*, as I've written about here before. In addition to that, there were a number of small items that I'm continuing to work on.

First, I've become quite fond of the question "*Does future me have a comparative advantage?*" Especially for small items, if the answer is "No" (and it's no far more often than it's yes) then just do it right now. The more trivial the task, the more useful it is. For instance, today I asked myself that while standing in the bedroom wondering whether to take 30 seconds to move my ExOfficio Bugproof socks from the dresser to the correct box in the closet. (Answer from a few minutes ago: if I don't take my dog for a walk *right now*, he's going to pee all over the floor. Future me does have a comparative advantage of not having to clean up pee on the floor. The socks can wait.)

I've begun to *notice my confusion and call it to conscious attention* more often, though I suspect I learned this first from HpMOR and the sequences before the workshop. Example: when Leonard Susskind states that conservation of information is a fundamental principle of quantum mechanics, I notice that I am confused because A) I have never heard of any such fundamental law of physics as information conservation B) Every definition of information I have ever heard indicates that information most certainly can be destroyed. So just what the heck is he talking about anyway? I am now making a conscious effort to research this topic rather than letting it slide by.

The workshop introduced me to the concepts of System 1 and System 2. System 1 is the faster, reactive, intuitive mind that uses heuristics and experience to react quickly. System 2 is the slower, analytical, logical, mathematical mind. I didn't immediately grok this or see how to apply it. However the workshop did convince me to read Daniel Kahneman's Thinking Fast and Slow, and I'm beginning to follow this. It could be useful going forward. I particularly like the examples given at the end of each chapter.

Similarly I completely did not understand the concepts of inside view vs. outside view at the workshop; and worse yet I don't think that I even realized that I didn't understand these. However now that I've read Thinking Fast and Slow, the lightbulb has gone on. Inside view is simply me deciding how likely I (or my team) is likely to accomplish something based on my judgement of the problem and our capabilities. Outside view is a statistical question about how people and teams like us have done when confronted with similar problems in the past. As long as there are similar teams and similar problems to compare with, the outside view is likely to be much more accurate.

During conversation, Julia Galef and I came up with the idea of *********. It turned out it already exists, and I'm planning to start attending these events locally soon. I've also joined my local LessWrong meetup group.

Stare into Ugh fields. Difficult conversations are an Ugh field for me. Recognizing this and bringing it to conscious attention has made it somewhat easier to manage these conversations. Example: when I went to the workshop I had been putting off contacting my dentist for months, not because of the usual reasons people don't like going to the dentist, but simply because I was uncomfortable telling her that the second (and third) opinion I had gotten on a dental issue disagreed with her about the proper course of treatment. Post-workshop, I finally called her (though it still took me two more weeks to do this. Clearly I have a lot of work left to do here.)

Consider whether the sources of my information may be correlated and by how much. I.e. Evaluating Advice. For instance, if two dentists who share an office give me the same advice, even assuming no prior disposition to agree with each other simply out of friendship, how likely is it that they share the same background and information that dentists in a different office do not?

COZE (Comfort Zone Expansion) exercises have pushed me to talk more to "strangers" and be intentionally more extroverted. On a recent trip to Latin America, I even made an effort to use what little Spanish I possess. I've had some small success, though this has led to no obvious major improvements in my life yet.

Thought experiments conducted at the workshop were very helpful in untangling some of my goals and plans. Going forward though this hasn't made a huge difference in my day-to-day life. That is, it hasn't led me to seek different paths than what I'm on right now.

What I Learned and Forgot

Going over my notes now, there was a lot of material; some of it potentially useful, that has fallen by the wayside; and may be worth a second look. This includes:

- Geoff Anders introduced us to yEd, a nice open source diagram editor. I still prefer StencilIt or Omnigraffle though. He also used it to show us a really neat way of graphing, well, something. Goals maybe? I remember it seemed really useful and significant at the time, but for the life of me I can't remember exactly what it was or what it was supposed to show us. I'll have to go back to my notes. This is why we write things down. (Update: I suspect this was about *Goal Factoring*.)
- Anticipation vs. Profession (though from time to time I do find myself asking what odds I'd be willing to bet on certain beliefs)
- The Planning Kata.

What I Learned But Didn't Implement

Value of Information calculations seem too meta and too wishy-washy to be of much use. They attempt to put quantitative numbers based on information that's far too imprecise to allow even order of magnitude accuracy. I'm better off just keeping things I need to consider in my GTD system, and periodically reviewing it.

Similarly opportunities for Bayesian Strength of Evidence calculations, just don't seem to come up in my day-to-day life. The question for me is more commonly "Given that the situation is what it is, what actions should I take to accomplish my goals?" The outside view is useful for this. Figuring out why the situation is what it is rarely seems to be especially helpful.

Turbocharging Training may be helpful but the evidence seems to me to be lacking. I'd like to see some strong proof that this works in particular areas; e.g. foreign languages, sports, or mathematics. Furthermore, it's not clear that it's applicable to anything I'm working on learning at this time. It seems very System 1 focused, and not especially helpful with the sort of fundamentally System 2 tasks I take on.

I have begun to declare "Victory!" at the end of a meeting/discussion. it's a bit of fun, but has limited effect. Beyond that I don't seem to reward myself for noticing things, or as a means of installing habits.

What I Didn't Learn

Getting Things Done (GTD), Remember the Milk, BeeMinder, Anki, Cultivating Curiosity, Overcoming Procrastination, and Winning at Arguments.

GTD I didn't learn because I've used it for years now or at least the parts of it that really work for me (lists and calendars mostly, and to a lesser extent filing).

Remember the Milk because my employer's security policy prohibits us from using it, and too much of my life happens at my day job to make maintaining two separate systems worthwhile.

BeeMinder and Anki because I just don't have anything that seems it could benefit from being stored in those systems right now. All of these might be more beneficial to someone in different circumstances.

Cultivating Curiosity because I am already a very naturally curious person, and have been for as long as I can remember. I don't need help with this. Indeed if anything I need to tamp down on this tendency and focus more on accomplishing things rather than merely learning them.

Similarly, *Overcoming Procrastination* didn't help a lot because I don't have a big procrastination problem, at least not compared to what I had when I was younger. Of course, I do say that in full knowledge that right this minute writing this article is a form of structured procrastination to avoid doing my taxes. :-)

Winning at Arguments, I am already very, very good at when I want to be, which is rare these days. It took me many years too realize that even though I "won" almost every argument I cared about, winning the argument wasn't usually all that useful. Winning an argument is the wrong goal to have for almost any purpose, and rarely leads to the outcomes I desire.

Unofficial ideas from fellow attendees:

Polyphasic sleep: I'm going to let the younger, more pioneering attendees experiment with this one. Even if it does work (which seems far from obvious) I don't see how one could integrate it into a conventional day job and family.

At breakfast one morning, a fellow attendee (Hunter?) suggested putting unsalted butter in my coffee to add more fat to my diet. It's not as crazy as it sounds. After all butter is little more than clarified cream, which I do like in my coffee. I tried this once and I still prefer cream, but I may give it another shot.

Finally, I've referred two workshop attendees to my employer as potential hires. If anyone else from the workshop is looking for a job, especially in tech, sales, or legal, drop me a line privately. For that matter if any Less Wronger is looking for a job, drop me a line privately. We have hundreds of open positions in major cities around the world. Quite a few LessWrongers already work there, and there's room for many more.

What the workshop didn't teach

There were a few techniques that were conspicuous by their absence. In particular I think the CFAR/LessWrong and Agile/XP communities have a lot to teach each other. I was surprised that no one at the workshop seemed to have heard of Kanban or Scrum, much less practice it. Burndown charts and point-based estimation are a really interesting modification of the outside view by comparing your team to your team in the past, rather than to other teams.

Pairing is also a useful technique beyond programming as at least Eliezer (not present at the workshop) has discovered. Pairing is an incredibly effective way to overcome akrasia and procrastination.

In reverse, I am considering what the craft of software development has to learn from CFAR style rationality, more specifically epistemic rationality. I have begun to notice my confusion during conversations with users, product managers, and tech leads and call it to conscious attention. I less frequently let unclear specs and goals pass without comment. Rather, I ask for examples and drill down into them until I feel my confusion has been conquered.

So far these techniques seem very useful in analysis and requirements gathering. I've found them less obviously useful (though certainly not harmful in any way) during coding, debugging, and testing. In these stages there's simply too much to be confused by to address it all, and whatever I'm confused by that's relevant to the task at hand rapidly calls itself to my attention. For instance, when a bug shows up in a production system, the very first and natural question to ask is "How the hell did the system do that?!" On the other hand, the planning kata may be very helpful with the early stages of system design, though I haven't yet had an opportunity to try that out.

Was it Worth $3900?

Overall, I found the workshop to be a worthwhile experience, if an expensive one; and I recommend it to you if you have the opportunity and resources to attend. There are a lot of practical techniques to be learned, and you only need one or two of them to pay off to cover the cost and time. Even if the primary value is simply introducing you to books and techniques you explore further after the workshop such as Getting Things Done or Thinking Fast and Slow, that may be enough. Most knowledge workers are operating far below the level of which we're capable, and expanding our effectiveness can pay for itself.

Before attending, it is worth asking yourself whether there's an opportunity to learn this material at lower cost. For instance, did I really need to spend $3900 and 4 days

to learn about Pomodoro? Apparently so, since I'd heard about Pomodoro for years and paid no attention to it until January. On the other hand, a $20 book I read on the subway was fully sufficient for me to learn and implement Getting Things Done. You'll have to judge this one for yourself.

We Don't Drink Vodka (LW Moscow report)

And we don't have bears playing balalaikas. Well, I would like to tell you about Moscow rationality community after all, not about some B movie featuring crazy Russians.

Moscow community have grown from 5 people on my first meetup to 17 on the last one. And I believe we have possibility to grow even more. Moscow is a big city and it must have many smart people who can start to study rationality.

Our story began in May 2012, when I gathered the people for the first time. Spring meetup announcements hadn't attracted many new members, so we gathered, discussed our site with Russian translations of LessWrong Sequences and made some plans. Our first venue was one of the Subway restaurants in Moscow.

The next milestone in our development was September meetup, when I started to use on-line form to collect information from potential members of our group. Or maybe for some other reason we had got new faces, and even recurring ones. I also told everyone than we should practice rationality skills doing some exercises. Of course we had a lot of theories and ideas to duscuss, but we had to be closer to the real world. That's why we started to practice our rationality skills. We have approximately 8-10 people on each meetup during this fall.

Soon enough this practice yielded good results, new members became heroes and started to improve our ways of training and create new exercises. In January, 2013 one member of our group proposed big and comfortable office room, and we moved there. Our meetups suddenly became more organized and more new members appeared — this year we have 13 people on average.

As for exercises, we practice calibration, tabooing, reframing, information value estimation and five minute debiasing. I will describe our versions of exercises below, and five minute debiasing you can find in the “How to Run a Successful Less Wrong Meetup” guidelines. Only a few people from our group can speak English well enough to come to CFAR workshops. I can't thank CFAR staff enough for their support for exercises adaptation. I hope they will be even more helpful in the future.

We have also started to design game that can teach group members some rationality skills. You can find some examples of interesting and fun games in the same guidelines I mentioned, but we want to develop games specially for the skills improvement. Of course even educational games have to produce fun, not only teach you something. We play Liar's dice for relaxation after exercises now.

And you can find some photos from our meetup here.

Appendix: Exercises

1. Calibration

Organizer presents two block of questions, each block has 10 questions for the sake of easy results calculation.

In the first one I read questions which require 90% confidence intervals. For example, what is the wingspan of the last model Boeing 737? It is similar to the one from "How to Run a Successful Less Wrong Meetup" guidelines. After this block everyone can calculate their real confidence, for example, if the correct answers are inside your interval in 7 out of 10, your confidence interval is closer to 70% than to 90%. So calibration level still can be improved in this case.

The second block is similar to The Credence Game, it consists of true or false questions. Everyone needs to write down credence for each answer. The average expected credence is calculated, then the real average is calculated. For example, if someone has the following credence: 0.5, 0.7, 0.8, 0.9, 0.7, 0.6, 0.6, 0.8, 0.9, 0.8 the average expected credence will be 0.73. And if there are 6 correct, 3 incorrect and one answer with credence 50%, then it will be 6.5 real average: 0.5 for one answer with 0.5 credence and 6 for the correct answers. The two numbers are close and the person with that answers seems to be well calibrated.

And for the some reason everyone showed better result in the second block. I can conclude that a person has more difficulties with hitting into specified confidence interval than assigning confidence for own answers.

During the calibration session I present the following strategies to improve calibration, once a time:

1. Repetition and feedback. Take several tests in succession, assessing how well you did after each one and attempting to improve your performance in the next one.
2. Equivalent bets. For each estimate, set up the equivalent bet to test if that range or probability really reflects your uncertainty. It means that you should choose between two games. In game A you will receive a money prize if your statement is true, in other words if the correct number is between your upper and lower bounds. In game B you generate random number between 0 and 1, and you win if the random is between zero and your credence (0.9, for example). I can say that you win with probability equals to your credence. If you prefer game A, you may be underconfident; if you prefer game B, you may be overconfident.
3. Consider two pros and two cons. Think of at least two reasons why you should be confident in your assessment and two reasons you could be wrong.
4. Avoid anchoring. Think of range questions as two separate binary questions of the form "Are you 95% certain that the true value is over/under (pick one) the lower/upper (pick one) bound?"
5. Reverse the anchoring effect. Start with extremely wide ranges and narrow them with the "absurdity test" as you eliminate highly unlikely values.

I recommend to make at least one calibration session before any Fermi calculation sessions.

2. Tabooing, version 2

There is standard rationalists' taboo exercise, you just remove some word from your speech and try to talk about something. But I would like to propose another version.

You need to create some texts, or use existing texts from a book or a news article. You also need to find some words in each text that should obscure the meaning therefore tabooing will make it clearer. It may be some texts about politics or other controversial

topic. Each text should be short, one or two paragraphs. You need to highlight words to taboo somehow, with italic font, separate colour or highlighter.

At the meetup ask the people to read the text tabooing the words you have selected.

If you are going to try this exercise, please let me know about the results, because I am still trying to improve it.

3. Reframing

First, define or find a statement you want to work with. The statement can be associated with some choice you want to make or it can be your interlocutor's phrase you want to make clear.

Second, do the reframing itself:

Check for you desire to maintain status quo. Do you see changes as bad things? Try to reverse changes direction.

Imagine, that you make a decision for every similar situation in the future.

Change unit of measurement, for example, convert time into money or vice versa.

Change time frame, into the past or the future.

Change the arena. Transfer your conditions into another country, even into imaginary place described in some book.

Imagine, that another person is faced with you issue. What will he or she decide or do?

Imagine yourself as an outside observer. What will you think about your own thoughts and deeds?

4. Value of information

Information can have an influence upon the utility and yield of your choices. If it can help you to make the choice with higher utility, then the difference is the value of the information for you. Take several daily situations when you need to make simple choices and estimate the impact of new information on these choices. For example, you need to buy some things and you may make random choice or look for detailed descriptions and specifications of this things. How much money, time and other resources you can save or earn in the future if you make informative choice instead of choice without the information?

Suggestion: Read Paul Graham

This isn't really a full post, but merely a note of potential interest. Paul Graham (who runs Hacker News) has several very interesting and thought-provoking essays located on his personal website. To me they fit very well with the style of thinking employed and advocated by many people on LW and I'd advise that nearly anyone interested in LW check out his work.

I especially recommend Keep Your Identity Small, What You Can't Say, and What You'll Wish You'd Known, but nearly every essay up there is interesting to me in some way. Many of them are directly relevant to issues of rationality, while others are only indirectly related, but either way I found them worth my time.

Don't Get Offended

Related to: Politics is the Mind-Killer, Keep Your Identity Small

Followed By: How to Not Get Offended

One oft-underestimated threat to epistemic rationality is getting offended. While getting offended by something sometimes feels good and can help you assert moral superiority, in most cases it doesn't help you figure out what the world looks like. In fact, getting offended usually makes it *harder* to figure out what the world looks like, since it means you won't be evaluating evidence very well. In Politics is the Mind-Killer, Eliezer writes that "people who would be level-headed about evenhandedly weighing all sides of an issue in their professional life as scientists, can suddenly turn into slogan-chanting zombies when there's a Blue or Green position on an issue." Don't let yourself become one of those zombies-- all of your skills, training, and useful habits can be shut down when your brain kicks into offended mode!

One might point out that getting offended is a two-way street and that it might be more appropriate to make a post called "Don't Be Offensive." That feels like a just thing to say-- as if you are targeting the aggressor rather than the victim. And on a certain level, it's true-- you *shouldn't* try to offend people, and if you do in the course of a normal conversation it's probably your fault. But you can't always rely on others around you being able to avoid doing this. After all, what's offensive to one person may not be so to another, and they may end up offending you by mistake. And even in those unpleasant cases when you are interacting with people who are deliberately trying to offend you, isn't staying calm desirable anyway?

The other problem I have with the concept of being offended as victimization is that, when you find yourself getting offended, you may be a victim, but you're being victimized by *yourself.* Again, that's not to say that offending people on purpose is acceptable-- it obviously isn't. But you're the one who gets to decide whether or not to be offended by something. If you find yourself getting offended to things as an automatic reaction, you should seriously evaluate why that is your response.

There is nothing inherent in a set of words that makes them offensive or inoffensive-- your reaction is an internal, personal process. I've seen some people stay cool in the face of others literally screaming racial slurs in their faces and I've seen other people get offended by the slightest implication or slip of the tongue. What type of reaction you have is largely up to you, and if you don't like your current reactions you can train better ones-- this is a core principle of the extremely useful philosophy known as Stoicism.

Of course, one (perhaps Robin Hanson) might also point out that getting offended can be socially useful. While true-- quickly responding in an offended fashion can be a strong signal of your commitment to group identity and values[1]-- that doesn't really relate to what this post is talking about. This post is talking about the best way to acquire correct beliefs, not the best way to manipulate people. And while getting offended can be a very effective way to manipulate people-- and hence a tactic that is unfortunately often reinforced-- it is usually actively detrimental for acquiring correct beliefs. Besides, the signalling value of offense should be no excuse for not knowing how not to be offended. After all, if you find it socially necessary to pretend that you are offended, doing so is not exactly difficult.

Personally, I have found that the cognitive effort required to build a habit of not getting offended pays immense dividends. Getting offended tends to shut down other mental processes and constrain you in ways that are often undesirable. In many situations, misunderstandings and arguments can be diminished or avoided completely if one is unwilling to become offended and practiced in the art of avoiding offense. Further, some of those situations are ones in which thinking clearly is very important indeed! All in all, while getting offended does often feel good (in a certain crude way), it is a reaction that I have no regrets about relinquishing.

[1] In Keep Your Identity Small, Paul Graham rightly points out that one way to prevent yourself from getting offended is to let as few things into your identity as possible.

Programming the LW Study Hall

We've had considerable interest and uptake on the Less Wrong Study Hall, especially with informal timed Pomodoro sessions for everyone to synchronize on. Working together with a number of other visible faces, and your own face visible to them, does seem effective. Keeping the social chat to the 5 off minutes prevents this from turning into just another chatroom.

We've been using this Tinychat room, and implementing everything Pomodoro-related with manual typing. Is there anyone out there who's interested in taking this to the next level with some custom code, possibly via the Google Hangouts API (Javascript), so we can have the following nice features?

- Synchronized, software-implemented Pomodoros for everyone.
    - Maybe one chat room with 20/5 and one with 45/5.
    - Actual enforcement of the "no chatting unless the Pomodoro is on" and/or muted microphones.
- Chat rooms with an N-person limit (several people report being more productive in smaller groups, so we're not sure if N should be 5 or 10) and new chat rooms being spawned as earlier ones get full.
- Moderatability (we've already had one troll). One person suggested the ability to +1/-1 one person per day, with a kick at -5. (Eliezer remarks that he expects this to be completely ineffective and that you need actual mods, maybe a group of trusted users.)
    - We only wish and dream that we could integrate with LW logins, but this would require LW development resources that apparently don't exist. Maybe with enough grotesque hackery we could have a page somewhere that you comment to confirm you're a Study Hall user, and the system could look up your karma from there to decide if you can cast +1/-1 user votes.
- A custom page layout (which you don't get if you use Tinychat!) which has the watching participants lined up vertically on the left, so we can work on something while still easily seeing our friends.
    - A small line underneath everyone's image saying what they're currently working on.
- Maybe a common room where people can initially talk about what they intend to work on. (Eliezer says: This needs either strong group norms or built-in limits on talk time to avoid becoming a social chat timesink.)
- The ability to branch off small sub-chat rooms (maybe with limit 2 or 3) in case somebody wants to talk (about work!)
- A welcome page (mockup) where people see the group norms the first time they visit the Study Hall that serves as a portal.

This doesn't "seem" very complicated from a programming perspective (yes, we all know about things that don't seem complicated). The Google Hangouts API (possibly OpenMeetings) seems like it should provide almost all of the basics already. But unless some particular programmer steps up to do it, it won't get done. If interested, comment below or email shannon.friedman@positivevector.com, and please mention your relevant Javascript experience.