Best of LessWrong: August 2013

Best of LessWrong: August 2013

Update on establishment of Cambridge's Centre for Study of Existential Risk

Cambridge's high-profile launch of the Centre for Study of Existential Risk last November received a lot of attention on LessWrong, and a number of people have been enquiring as to what's happened since. This post is meant to give a little explanation and update of what's been going on.

Motivated by a common concern over human activity-related risks to humanity, Lord Martin Rees, Professor Huw Price, and Jaan Tallinn founded the Centre for Study of Existential Risk last year. However, this announcement was made before the establishment of a physical research centre or securement of long-term funding. The last 9 months have been focused on turning an important idea into a reality.

Following the announcement in November, Professor Price contacted us at the Future of Humanity Institute regarding the possibility of collaboration on joint academic funding opportunities; the aim being both to raise the funds for CSER's research programmes and to support joint work by the FHI and CSER's researchers on anthropogenic existential risk. We submitted our first grant application in January to the European Research Council – an ambitious project to create "A New Science of Existential Risk" that, if successful, would provide enough funding for CSER's first research programme - a sizeable programme that will run for five years.
We've been successful in the first and second rounds, and we will hear a final round decision at the end of the year. It was also an opportunity for us to get some additional leading academics onto the project – Sir Partha Dasgupta, Professor of Economics at Cambridge and an expert in social choice theory, sustainability and intergenerational ethics, is a co-PI (along with Huw Price, Martin Rees and Nick Bostrom). In addition, a number of prominent academics concerned about technology-related risk – including Stephen Hawking, David Spiegelhalter, George Church and David Chalmers – have joined our advisory board.

The FHI regards establishment of CSER as of the highest priority for a number of reasons including:

1) The value of the research the Centre will engage in
2) The reputational boost to the field of Existential Risk gained by the establishment of high-profile research centre in Cambridge.
3) The impact on policy and public perception that academic heavy-hitters like Rees and Price can have

Therefore we've been working with CSER behind the scenes over the last 9 months. Progress has been a little slow until now – Huw, Martin and Jaan are fully committed to this project, but due to their other responsibilities aren't in a position to work full-time on it yet.

However, we're now in a position to make CSER's establishment official. Cambridge's new Centre for Research in the Arts, Social Sciences and Humanities (CRASSH) will host CSER and provide logistical support. I'll be acting manager of CSER's activities over the coming 6-12 months, under the guidance of Huw, Martin and Jaan. A generous seed funding donation from Jaan Tallinn is funding CSER's establishment and these activities – which will include a lecture series, workshops, public outreach, and staff time on grant-writing and fundraising. It'll also provide a buyout of a fraction of my time from FHI (providing funds for us to hire part-time staff to offload some of the FHI workload and help with some of the CSER work).

At the moment and over the next couple of months we're going to be focused on identifying and working on additional academic funding opportunities for additional programmes, as well as chasing some promising leads in industry, private and philanthropic funding. I'll also be aiming to keep CSER's public profile active. There will be newsletters every three months (sign up here), the website's going to be fleshed out to contain more detail about our planned research and existing literature, and we'll be arranging regular high-quality media engagement. While we're unlikely to have time to answer every general query that comes in (though we'll try whenever possible: email: admin@cser.org), we'll aim to keep the existential risk community informed through the newsletters and posts such as these.

We've been lucky to get a lot of support from the academic and existential risk community for the CSER centre. In addition to CRASSH, Cambridge's Centre for Science and Policy will provide support in making policy-relevant links, and may co-host and co-publicise events. Luke Muehlhauser, MIRI's Executive Director, has been very supportive and has provided valuable advice, and has generously offered to direct some of MIRI's volunteer support towards CSER tasks. We also expect to get valuable support from the growing community around FHI.

From where I'm sitting, CSER's successful launch is looking very promising. The timeline on our research programmes, however, is still a little more uncertain. If we're successful with the European Research Council, we can expect to be hiring a full research team next spring. If not, it may take a little longer, but we're exploring a number of different opportunities in parallel and are feeling confident. The support of the existential risk community continues to be invaluable.

Thanks,

Seán Ó hÉigeartaigh
Academic Manager, Future of Humanity Institute

http://www.fhi.ox.ac.uk

Acting Academic Manager, Cambridge Centre for Study of Existential Risk.

http://www.cser.org

Use Your Identity Carefully

In Keep Your Identity Small, Paul Graham argues against associating yourself with labels (i.e. “libertarian,” “feminist,” “gamer,” “American”) because labels constrain what you’ll let yourself believe. It’s a wonderful essay that’s led me to make concrete changes in my life. That said, it’s only about 90% correct. I have two issues with Graham’s argument; one is a semantic quibble, but it leads into the bigger issue, which is a tactic I’ve used to become a better person.

Graham talks about the importance of identity in determining beliefs. This isn’t quite the right framework. I’m a fanatical consequentialist, so I care what actions people take. Beliefs can constrain actions, but identity can also constrain actions directly.

To give a trivial example from the past week in which beliefs didn’t matter: I had a self-image as someone who didn’t wear jeans or t-shirts. As it happens, there are times when wearing jeans is completely fine, and when other people wore jeans in casual settings, I knew it was appropriate. Nevertheless, I wasn’t able to act on this belief because of my identity. (I finally realized this was silly, consciously discarded that useless bit of identity, and made a point of wearing jeans to a social event.)

Why is this distinction important? If we’re looking at identify from an action-centered framework, this recommends a different approach from Graham’s.

Do you want to constrain your beliefs? No; you want to go wherever the evidence pushes you. “If X is true, I desire to believe that X is true. If X is not true, I desire to believe that X is not true.” Identity will only get in the way.

Do you want to constrain your actions? Yes! Ten thousand times yes! Akrasia exists. Commitment devices are useful. Beeminder is successful. Identity is one of the most effective tools for the job, if you wield it deliberately.

I’ve cultivated an identity as a person who makes events happen. It took months to instill, but now, when I think “I wish people were doing X,” I instinctively start putting together a group to do X. This manifests in minor ways, like the tree-climbing expedition I put together at the Effective Altruism Summit, and in big ways, like the megameetup we held in Boston. If I hadn’t used my identity to motivate myself, neither of those things would’ve happened, and my life would be poorer.

Identity is powerful. Powerful things are dangerous, like backhoes and bandsaws. People use them anyway, because sometimes they’re the best tools for the job, and because safety precautions can minimize the danger.

Identity is hard to change. Identity can be difficult to notice. Identity has unintended consequences. Use this tool only after careful deliberation. What would this identity do to your actions? What would it do to your beliefs? What social consequences would it have? Can you do the same thing with a less dangerous tool? Think twice, and then think again, before you add to your identity. Most identities are a hindrance.

But please, don’t discard this tool just because some things might go wrong. If you are willful, and careful, and wise, then you can cultivate the identity of the person you always wanted to be.

What Bayesianism taught me

David Chapman criticizes "pop Bayesianism" as just common-sense rationality dressed up as intimidating math[1]:

> Bayesianism boils down to “don’t be so sure of your beliefs; be less sure when you see contradictory evidence.”
>
> Now that is just common sense. Why does anyone need to be told this? And how does [Bayes'] formula help?
>
> [...]
>
> The leaders of the movement presumably do understand probability. But I’m wondering whether they simply use Bayes’ formula to intimidate lesser minds into accepting “don’t be so sure of your beliefs.” (In which case, Bayesianism is not about Bayes’ Rule, after all.)
>
> I don’t think I’d approve of that. “Don’t be so sure” is a valuable lesson, but I’d rather teach it in a way people can understand, rather than by invoking a Holy Mystery.

What does Bayes's formula have to teach us about how to do epistemology, beyond obvious things like "never be absolutely certain; update your credences when you see new evidence"?

I list below some of the specific things that I learned from Bayesianism. Some of these are examples of mistakes I'd made that Bayesianism corrected. Others are things that I just hadn't thought about explicitly before encountering Bayesianism, but which now seem important to me.

I'm interested in hearing what other people here would put on their own lists of things Bayesianism taught them. (Different people would make different lists, depending on how they had already thought about epistemology when they first encountered "pop Bayesianism".)

I'm interested especially in those lessons that you think followed more-or-less directly from taking Bayesianism seriously as a normative epistemology (plus maybe the idea of making decisions based on expected utility). The LW memeplex contains many other valuable lessons (e.g., avoid the mind-projection fallacy, be mindful of inferential gaps, the MW interpretation of QM has a lot going for it, decision theory should take into account "logical causation", etc.). However, these seem further afield or more speculative than what I think of as "bare-bones Bayesianism".

So, without further ado, here are some things that Bayesianism taught me.

1. **Banish talk like "There is absolutely no evidence for that belief".** $P(E \mid H) > P(E)$ if and only if $P(H \mid E) > P(H)$. The fact that there are myths about Zeus is evidence that Zeus exists. Zeus's existing would make it more likely for myths about him to arise, so the arising of myths about him must make it more likely that he exists. A related mistake I made was to be impressed by the cleverness of the aphorism "The plural of 'anecdote' is not 'data'." There may be a helpful

distinction between scientific evidence and Bayesian evidence. But anecdotal evidence is evidence, and it ought to sway my beliefs.

2. **Banish talk like "I don't know anything about that".** See the post "I don't know."
3. **Banish talk of "thresholds of belief".** Probabilities go up or down, but there is no magic threshold beyond which they change qualitatively into "knowledge". I used to make the mistake of saying things like, "I'm not absolutely certain that atheism is true, but it is my working hypothesis. I'm confident enough to act as though it's true." I assign a certain probability to atheism, which is less than 1.0. I ought to act as though I am just that confident, and no more. I should never *just assume* that I am in the possible world that I think is most likely, even if I think that that possible world is *overwhelmingly* likely. (However, perhaps I could be so confident that my behavior would not be practically discernible from absolute confidence.)
4. **Absence of evidence *is* evidence of absence**. $P(H \mid E) > P(H)$ if and only if $P(H \mid {\sim}E) < P(H)$. Absence of evidence may be very weak evidence of absence, but it is evidence nonetheless. (However, you may not be entitled to a particular kind of evidence.)
5. **Many bits of "common sense" rationality can be *precisely stated* and *easily proved* within the austere framework of Bayesian probability.** As noted by Jaynes in *Probability Theory: The Logic of Science*, "[P]robability theory as extended logic reproduces many aspects of human mental activity, sometimes in surprising and even disturbing detail." While these things might be "common knowledge", the fact that they are readily deducible from a few simple premises is significant. Here are some examples:
   - **It is possible for the opinions of different people to *diverge* after they *rationally* update on the same evidence.** Jaynes discusses this phenomenon in Section 5.3 of *PT:TLoS*.
   - **Popper's falsification criterion**, and other Popperian principles of "good explanation", such as that good explanations should be "hard to vary", follow from Bayes's formula. Eliezer discusses this in *An Intuitive Explanation of Bayes' Theorem* and *A Technical Explanation of Technical Explanation*.
   - **Occam's razor**. This can be formalized using Solomonoff induction. (However, perhaps this shouldn't be on my list, because Solomonoff induction goes beyond just Bayes's formula. It also has several problems.)
6. **You cannot expect[2] that future evidence will sway you in a particular direction.** "For every expectation of evidence, there is an equal and opposite expectation of counterevidence."
7. **Abandon all the meta-epistemological intuitions about the concept of knowledge on which Gettier-style paradoxes rely.** Keep track of how confident your beliefs are when you update on the evidence. Keep track of the extent to which other people's beliefs are good evidence for what they believe. Don't worry about whether, in addition, these beliefs qualify as "knowledge".

What items would you put on your list?

ETA: ChrisHallquist's post Bayesianism for Humans lists other "directly applicable corollaries to Bayesianism".

[1] See also Yvain's reaction to David Chapman's criticisms.

[2] **ETA:** My wording here is potentially misleading. See this comment thread.

Humans are utility monsters

When someone complains that utilitarianism[1] leads to the dust speck paradox or the trolley-car problem, I tell them that's a feature, not a bug. I'm not ready to say that respecting the utility monster is also a feature of utilitarianism, but it *is* what most people everywhere have always done. A model that doesn't allow for utility monsters can't model human behavior, and certainly shouldn't provoke indignant responses from philosophers who keep right on respecting their own utility monsters.

The utility monster is a creature that is somehow more capable of experiencing pleasure (or positive utility) than all others combined. Most people consider sacrificing everyone else's small utilities for the benefits of this monster to be repugnant.

Let's suppose the utility monster is a utility monster because it has a more highly-developed brain capable of making finer discriminations, higher-level abstractions, and more associations than all the lesser minds around it. Does that make it less repugnant? (If so, I lose you here. I invite you to post a comment explaining why utility-monster-by-smartness is an exception.) Suppose we have one utility monster and one million others. Everything we do, we do for the one utility monster. Repugnant?

Multiply by nine billion. We now have nine billion utility monsters and 9x10^{15} others. Still repugnant?

Yet these same enlightened, democratic societies whose philosophers decry the utility monster give approximately zero weight to the well-being of non-humans. We might try not to drive a species *extinct*, but when contemplating a new hydroelectric dam, nobody adds up the disutility to all the squirrels in the valley to be flooded.

If you believe the utility monster is a problem with utilitarianism, how do you take into account the well-being of squirrels? How about ants? Worms? Bacteria? You've gone to 10^{15} others just with ants.[2] Maybe 10^{20} with nematodes.

"But humans are different!" our anti-utilitarian complains. "They're so much more intelligent and emotionally complex than nematodes that it would be repugnant to wipe out all humans to save any number of nematodes."

Well, that's what a real utility monster looks like.

The same people who believe this then turn around and say there's a problem with utilitarianism because (when unpacked into a plausible real-life example) it might kill all the nematodes to save one human. Given their beliefs, they should complain about the opposite "problem": For a sufficient number of nematodes, an instantiation of utilitarianism might say *not* to kill all the nematodes to save one human.

[1] I use the term in a very general way, meaning any action selection system that uses a utility function—which in practice means any rational, deterministic action selection system in which action preferences are well-ordered.

[2] This recent attempt to estimate the number of different living beings of different kinds gives some numbers. The web has many pages claiming there are 10^{15} ants, but

I haven't found a citation of any original source.

How to Measure Anything

Douglas Hubbard's *How to Measure Anything* is one of my favorite how-to books. I hope this summary inspires you to buy the book; it's worth it.

The book opens:

> Anything can be measured. If a thing can be observed in any way at all, it lends itself to some type of measurement method. No matter how "fuzzy" the measurement is, it's still a measurement if it tells you more than you knew before. And those very things most likely to be seen as immeasurable are, virtually always, solved by relatively simple measurement methods.

The sciences have many established measurement methods, so Hubbard's book focuses on the measurement of "business intangibles" that are important for decision-making but tricky to measure: things like management effectiveness, the "flexibility" to create new products, the risk of bankruptcy, and public image.

Basic Ideas

A *measurement* is an observation that quantitatively reduces uncertainty. Measurements might not yield precise, certain judgments, but they *do* reduce your uncertainty.

To be measured, the *object of measurement* must be described clearly, in terms of observables. A good way to clarify a vague object of measurement like "IT security" is to ask "What is IT security, and why do you care?" Such probing can reveal that "IT security" means things like a reduction in unauthorized intrusions and malware attacks, which the IT department cares about because these things result in lost productivity, fraud losses, and legal liabilities.

Uncertainty is the lack of certainty: the true outcome/state/value is not known.

Risk is a state of uncertainty in which some of the possibilities involve a loss.

Much pessimism about measurement comes from a lack of experience making measurements. Hubbard, who is *far* more experienced with measurement than his readers, says:

1. Your problem is not as unique as you think.
2. You have more data than you think.
3. You need less data than you think.
4. An adequate amount of new data is more accessible than you think.

Applied Information Economics

Hubbard calls his method "Applied Information Economics" (AIE). It consists of 5 steps:

1. Define a decision problem and the relevant variables. (Start with the decision you need to make, then figure out which variables would make your decision easier if you had better estimates of their values.)
2. Determine what you know. (Quantify your uncertainty about those variables in terms of ranges and probabilities.)
3. Pick a variable, and compute the value of additional information for that variable. (Repeat until you find a variable with reasonably high information value. If no remaining variables have enough information value to justify the cost of measuring them, skip to step 5.)
4. Apply the relevant measurement instrument(s) to the high-information-value variable. (Then go back to step 3.)
5. Make a decision and act on it. (When you've done as much uncertainty reduction as is economically justified, it's time to act!)

These steps are elaborated below.

Step 1: Define a decision problem and the relevant variables

Hubbard illustrates this step by telling the story of how he helped the Department of Veterans Affairs (VA) with a measurement problem.

The VA was considering seven proposed IT security projects. They wanted to know "which… of the proposed investments were justified and, after they were implemented, whether improvements in security justified further investment…" Hubbard asked his standard questions: "What do you mean by 'IT security'? Why does it matter to you? What are you observing when you observe improved IT security?"

It became clear that *nobody* at the VA had thought about the details of what "IT security" meant to them. But after Hubbard's probing, it became clear that by "IT security" they meant a reduction in the frequency and severity of some undesirable events: agency-wide virus attacks, unauthorized system access (external or internal),unauthorized physical access, and disasters affecting the IT infrastructure (fire, flood, etc.) And each undesirable event was on the list because of specific costs associated with it: productivity losses from virus attacks, legal liability from unauthorized system access, etc.

Now that the VA knew what they meant by "IT security," they could measure specific variables, such as the number of virus attacks per year.

Step 2: Determine what you know

Uncertainty and calibration

The next step is to determine your level of uncertainty about the variables you want to measure. To do this, you can express a "confidence interval" (CI). A 90% CI is a range of values that is 90% likely to contain the correct value. For example, the security experts at the VA were 90% confident that each agency-wide virus attack would affect between 25,000 and 65,000 people.

Unfortunately, few people are well-calibrated estimators. For example in some studies, the true value lay in subjects' 90% CIs only 50% of the time! These subjects were overconfident. For a well-calibrated estimator, the true value will lie in her 90% CI roughly 90% of the time.

Luckily, "assessing uncertainty is a general skill that can be taught with a measurable improvement."

Hubbard uses several methods to calibrate each client's value estimators, for example the security experts at the VA who needed to estimate the frequency of security breaches and their likely costs.

His first technique is the *equivalent bet test*. Suppose you're asked to give a 90% CI for the year in which Newton published the universal laws of gravitation, and you can win $1,000 in one of two ways:

1. You win $1,000 if the true year of publication falls within your 90% CI. Otherwise, you win nothing.
2. You spin a dial divided into two "pie slices," one covering 10% of the dial, and the other covering 90%. If the dial lands on the small slice, you win nothing. If it lands on the big slice, you win $1,000.

If you find yourself preferring option #2, then you must think spinning the dial has a higher chance of winning you $1,000 than option #1. That suggest your stated 90% CI isn't really your 90% CI. Maybe it's your 65% CI or your 80% CI instead. By preferring option #2, your brain is trying to tell you that your originally stated 90% CI is overconfident.

If instead you find yourself preferring option #1, then you must think there is *more* than a 90% chance your stated 90% CI contains the true value. By preferring option #1, your brain is trying to tell you that your original 90% CI is under confident.

To make a better estimate, adjust your 90% CI until option #1 and option #2 seem equally good to you. Research suggests that even *pretending* to bet money in this way will improve your calibration.

Hubbard's second method for improving calibration is simply *repetition and feedback*. Make lots of estimates and then see how well you did. For this, play CFAR's Calibration Game.

Hubbard also asks people to identify reasons why a particular estimate might be right, and why it might be wrong.

He also asks people to look more closely at each bound (upper and lower) on their estimated range. A 90% CI "means there is a 5% chance the true value could be greater than the upper bound, and a 5% chance it could be less than the lower bound. This means the estimators must be 95% sure that the true value is less than the upper bound. If they are not that certain, they should increase the upper bound… A similar test is applied to the lower bound."

Simulations

Once you determine what you know about the uncertainties involved, how can you use that information to determine what you know about the *risks* involved? Hubbard summarizes:

> …all risk in any project… can be expressed by one method: the ranges of uncertainty on the costs and benefits, and probabilities on events that might affect them.

The simplest tool for measuring such risks accurately is the Monte Carlo (MC) simulation, which can be run by Excel and many other programs. To illustrate this tool, suppose you are wondering whether to lease a new machine for one step in your manufacturing process.

> The one-year lease [for the machine] is $400,000 with no option for early cancellation. So if you aren't breaking even, you are still stuck with it for the rest of the year. You are considering signing the contract because you think the more advanced device will save some labor and raw materials and because you think the maintenance cost will be lower than the existing process.

Your pre-calibrated estimators give their 90% CIs for the following variables:

- Maintenance savings (MS): $10 to $20 per unit
- Labor savings (LS): -$2 to $8 per unit
- Raw materials savings (RMS): $3 to $9 per unit
- Production level (PL): 15,000 to 35,000 units per year

Thus, your annual savings will equal (MS + LS + RMS) × PL.

When measuring risk, we don't just want to know the "average" risk or benefit. We want to know the probability of a huge loss, the probability of a small loss, the probability of a huge savings, and so on. That's what Monte Carlo can tell us.

An MC simulation uses a computer to randomly generate thousands of possible values for each variable, based on the ranges we've estimated. The computer then calculates the outcome (in this case, the annual savings) for each generated combination of values, and we're able to see how often different kinds of outcomes occur.

To run an MC simulation we need not just the 90% CI for each variable but also the *shape* of each distribution. In many cases, the normal distribution will work just fine, and we'll use it for all the variables in this simplified illustration. (Hubbard's book shows you how to work with other distributions).

To make an MC simulation of a normally distributed variable in Excel, we use this formula:

```
=norminv(rand(), mean, standard deviation)
```

So the formula for the maintenance savings variable should be:

```
=norminv(rand(), 15, (20–10)/3.29)
```

Suppose you enter this formula on cell A1 in Excel. To generate (say) 10,000 values for the maintenance savings value, just (1) copy the contents of cell A1, (2) enter

“A1:A10000” in the cell range field to select cells A1 through A10000, and (3) paste the formula into all those cells.

Now we can follow this process in other columns for the other variables, including a column for the “total savings” formula. To see how many rows made a total savings of $400,000 or more (break-even), use Excel’s countif function. In this case, you should find that about 14% of the scenarios resulted in a savings of less than $400,000 – a loss.

We can also make a histogram (see right) to show how many of the 10,000 scenarios landed in each $100,000 increment (of total savings). This is even more informative, and tells us a great deal about the distribution of risk and benefits we might incur from investing in the new machine. (Download the full spreadsheet for this example here.)

The simulation concept can (and in high-value cases *should*) be carried beyond this simple MC simulation. The first step is to learn how to use a greater variety of distributions in MC simulations. The second step is to deal with correlated (rather than independent) variables by generating correlated random numbers or by modeling what the variables have in common.

A more complicated step is to use a Markov simulation, in which the simulated scenario is divided into many time intervals. This is often used to model stock prices, the weather, and complex manufacturing or construction projects. Another more complicated step is to use an agent-based model, in which independently-acting agents are simulated. This method is often used for traffic simulations, in which each vehicle is modeled as an agent.

Step 3: Pick a variable, and compute the value of additional information for that variable

Information can have three kinds of value:

1. Information can affect people’s behavior (e.g. common knowledge of germs affects sanitation behavior).
2. Information can have its own market value (e.g. you can sell a book with useful information).
3. Information can reduce uncertainty about important decisions. (This is what we’re focusing on here.)

When you’re uncertain about a decision, this means there’s a chance you’ll make a non-optimal choice. The cost of a “wrong” decision is the difference between the wrong choice and the choice you would have made with perfect information. But it’s too costly to acquire perfect information, so instead we’d like to know which decision-relevant variables are the *most* valuable to measure more precisely, so we can decide which measurements to make.

Here’s a simple example:

> Suppose you could make $40 million profit if [an advertisement] works and lose $5 million (the cost of the campaign) if it fails. Then suppose your calibrated

> experts say they would put a 40% chance of failure on the campaign.

The expected opportunity loss (EOL) for a choice is the probability of the choice being “wrong” times the cost of it being wrong. So for example the EOL if the campaign is approved is $5M × 40% = $2M, and the EOL if the campaign is rejected is $40M × 60% = $24M.

The difference between EOL before and after a measurement is called the “expected value of information” (EVI).

In most cases, we want to compute the VoI for a range of values rather than a binary succeed/fail. So let’s tweak the advertising campaign example and say that a calibrated marketing expert’s 90% CI for sales resulting from the campaign was from 100,000 units to 1 million units. The risk is that we don’t sell enough units from this campaign to break even.

Suppose we profit by $25 per unit sold, so we’d have to sell at least 200,000 units from the campaign to break even (on a $5M campaign). To begin, let’s calculate the expected value of *perfect* information (EVPI), which will provide an upper bound on how much we should spend to reduce our uncertainty about how many units will be sold as a result of the campaign. Here’s how we compute it:

1. Slice the distribution of our variable into thousands of small segments.
2. Compute the EOL for each segment. EOL = segment midpoint times segment probability.
3. Sum the products from step 2 for all segments.

Of course, we’ll do this with a computer. For the details, see Hubbard’s book and the Value of Information spreadsheet from his website.

In this case, the EVPI turns out to be about $337,000. This means that we shouldn’t spend more than $337,000 to reduce our uncertainty about how many units will be sold as a result of the campaign.

And in fact, we should probably spend much less than $337,000, because no measurement we make will give us *perfect* information. For more details on how to measure the value of *imperfect* information, see Hubbard’s book and these three LessWrong posts: (1) VoI: 8 Examples, (2) VoI: Four Examples, and (3) 5-second level case study: VoI.

I do, however, want to quote Hubbard’s comments about the “measurement inversion”:

> By 1999, I had completed the… Applied Information Economics analysis on about 20 major [IT] investments… Each of these business cases had 40 to 80 variables, such as initial development costs, adoption rate, productivity improvement, revenue growth, and so on. For each of these business cases, I ran a macro in Excel that computed the information value for each variable… [and] I began to see this pattern: * The vast majority of variables had an information value of zero… * The variables that had high information values were routinely those that the client had never measured… * The variables that clients [spent] the most time measuring were usually those with a very low (even zero) information value… … since then, I’ve applied this same test to another 40 projects, and… [I’ve] noticed the same phenomena arise in projects relating to research and development, military logistics, the environment, venture capital, and facilities expansion.

Hubbard calls this the "Measurement Inversion":

> In a business case, the economic value of measuring a variable is usually inversely proportional to how much measurement attention it usually gets.

Here is one example:

> A stark illustration of the Measurement Inversion for IT projects can be seen in a large UK-based insurance client of mine that was an avid user of a software complexity measurement method called "function points." This method was popular in the 1980s and 1990s as a basis of estimating the effort for large software development efforts. This organization had done a very good job of tracking initial estimates, function point estimates, and actual effort expended for over 300 IT projects. The estimation required three or four full-time persons as "certified" function point counters…
>
> But a very interesting pattern arose when I compared the function point estimates to the initial estimates provided by project managers… The costly, time-intensive function point counting did change the initial estimate but, on average, it was no closer to the actual project effort than the initial effort… Not only was this the single largest measurement effort in the IT organization, it literally added *no* value since it didn't reduce uncertainty at all. Certainly, more emphasis on measuring the benefits of the proposed projects – or almost anything else – would have been better money spent.

Hence the importance of calculating EVI.

Step 4: Apply the relevant measurement instrument(s) to the high-information-value variable

If you followed the first three steps, then you've defined a variable you want to measure in terms of the decision it affects and how you observe it, you've quantified your uncertainty about it, and you've calculated the value of gaining additional information about it. Now it's time to reduce your uncertainty about the variable – that is, to measure it.

Each scientific discipline has its own specialized measurement methods. Hubbard's book describes measurement methods that are often useful for reducing our uncertainty about the "softer" topics often encountered by decision-makers in business.

Selecting a measurement method

To figure out which category of measurement methods are appropriate for a particular case, we must ask several questions:

1. Decomposition: Which parts of the thing are we uncertain about?
2. Secondary research: How has the thing (or its parts) been measured by others?

3. Observation: How do the identified observables lend themselves to measurement?
4. Measure just enough: How much do we need to measure it?
5. Consider the error: How might our observations be misleading?

Decomposition

Sometimes you'll want to start by decomposing an uncertain variable into several parts to identify which observables you can most easily measure. For example, rather than directly estimating the cost of a large construction project, you could break it into parts and estimate the cost of each part of the project.

In Hubbard's experience, it's often the case that decomposition itself – even without making any new measurements – often reduces one's uncertainty about the variable of interest.

Secondary research

Don't reinvent the world. In almost all cases, someone has already invented the measurement tool you need, and you just need to find it. Here are Hubbard's tips on secondary research:

1. If you're new to a topic, start with Wikipedia rather than Google. Wikipedia will give you a more organized perspective on the topic at hand.
2. Use search terms often associated with quantitative data. E.g. don't just search for "software quality" or "customer perception" – add terms like "table," "survey," "control group," and "standard deviation."
3. Think of internet research in two levels: general search engines and topic-specific repositories (e.g. the CIA World Fact Book).
4. Try multiple search engines.
5. If you find marginally related research that doesn't directly address your topic of interest, check the bibliography more relevant reading material.

I'd also recommend my post Scholarship: How to Do It Efficiently.

Observation

If you're not sure how to measure your target variable's observables, ask these questions:

1. Does it leave a trail? Example: longer waits on customer support lines cause customers to hang up and not call back. Maybe you can also find a correlation between customers who hang up after long waits and reduced sales to those customers.
2. Can you observe it directly? Maybe you haven't been tracking how many of the customers in your parking lot show an out-of-state license, but you could start. Or at least, you can observe a sample of these data.

3. Can you create a way to observe it indirectly? Amazon.com added a gift-wrapping feature in part so they could better track how many books were being purchased as gifts. Another example is when consumers are given coupons so that retailers can see which newspapers their customers read.
4. Can the thing be forced to occur under new conditions which allow you to observe it more easily? E.g. you could implement a proposed returned-items policy in some stores but not others and compare the outcomes.

Measure just enough

Because initial measurements often tell you quite a lot, and also change the value of continued measurement, Hubbard often aims for spending 10% of the EVPI on a measurement, and sometimes as little as 2% (especially for very large projects).

Consider the error

It's important to be conscious of some common ways in which measurements can mislead.

Scientists distinguish two types of measurement error: systemic and random. Random errors are random variations from one observation to the next. They can't be individually predicted, but they fall into patterns that can be accounted for with the laws of probability. Systemic errors, in contrast, are consistent. For example, the sales staff may routinely overestimate the next quarter's revenue by 50% (on average).

We must also distinguish precision and accuracy. A "precise" measurement tool has low random error. E.g. if a bathroom scale gives the exact same displayed weight every time we set a particular book on it, then the scale has high precision. An "accurate" measurement tool has low systemic error. The bathroom scale, while precise, might be inaccurate if the weight displayed is systemically biased in one direction – say, eight pounds too heavy. A measurement tool can also have low precision but good accuracy, if it gives inconsistent measurements but they average to the true value.

Random error tends to be easier to handle. Consider this example:

> For example, to determine how much time sales reps spend in meetings with clients versus other administrative tasks, they might choose a complete review of all time sheets… [But] if a complete review of 5,000 time sheets… tells us that sales reps spend 34% of their time in direct communication with customers, we still don't know how far from the truth it might be. Still, this "exact" number seems reassuring to many managers. Now, suppose a sample of direct observations of randomly chosen sales reps at random points in time finds that sales reps were in client meetings or on client phone calls only 13 out of 100 of those instances. (We can compute this without interrupting a meeting by asking as soon as the rep is available.) As we will see [later], in the latter case, we can statistically compute a 90% CI to be 7.5% to 18.5%. Even though this random sampling approach gives us only a range, we should prefer its findings to the

> census audit of time sheets. The census… gives us an exact number, but we have no way to know by how much and in which direction the time sheets err.

Systemic error is also called a “bias.” Based on his experience, Hubbard suspects the three most important to avoid are:

1. Confirmation bias: people see what they want to see.
2. Selection bias: your sample might not be representative of the group you’re trying to measure.
3. Observer bias: the very act of observation can affect what you observe. E.g. in one study, researchers found that worker productivity improved no matter *what* they changed about the workplace. The workers seem to have been responding merely to the *fact* that they were being observed in *some* way.

Choose and design the measurement instrument

After following the above steps, Hubbard writes, “the measurement instrument should be almost completely formed in your mind.” But if you still can’t come up with a way to measure the target variable, here are some additional tips:

1. *Work through the consequences*. If the value is surprisingly high, or surprisingly low, what would you expect to see?
2. *Be iterative*. Start with just a few observations, and then recalculate the information value.
3. *Consider multiple approaches*. Your first measurement tool may not work well. Try others.
4. *What’s the really simple question that makes the rest of the measurement moot?* First see if you can detect *any* change in research quality before trying to measure it more comprehensively.

Sampling reality

In most cases, we’ll estimate the values in a population by measuring the values in a small sample from that population. And for reasons discussed in chapter 7, a very small sample can often offer large reductions in uncertainty.

There are a variety of tools we can use to build our estimates from small samples, and which one we should use often depends on how outliers are distributed in the population. In some cases, outliers are very close to the mean, and thus our estimate of the mean can converge quickly on the true mean as we look at new samples. In other cases, outliers can be several orders of magnitude away from the mean, and our estimate converges very slowly or not at all. Here are some examples:

- Very quick convergence, only 1–2 samples needed: cholesterol level of your blood, purity of public water supply, weight of jelly beans.
- Usually quickly convergence, 5–30 samples needed: Percentage of customers who like the new product, failure loads of bricks, age of your customers, how many movies people see in a year.

- Potentially slow convergence: Software project cost overruns, factory downtime due to an accident.
- Maybe non-convergent: Market value of corporations, individual levels of income, casualties of wars, size of volcanic eruptions.

Below, I survey just a few of the many sampling methods Hubbard covers in his book.

Mathless estimation

When working with a quickly converging phenomenon and a symmetric distribution (uniform, normal, camel-back, or bow-tie) for the population, you can use the t-statistic to develop a 90% CI even when working with very small samples. (See the book for instructions.)

Or, even easier, make use of the *Rule of FIve*: “There is a 93.75% chance that the median of a population is between the smallest and largest values in any random sample of five from that population.”

The Rule of Five has another advantage over the t-statistic: it works for any distribution of values in the population, including ones with slow convergence or no convergence at all! It can do this because it gives us a confidence interval for the *median* rather than the *mean*, and it’s the mean that is far more affected by outliers.

Hubbard calls this a “mathless” estimation technique because it doesn’t require us to take square roots or calculate standard deviation or anything like that. Moreover, this mathless technique extends beyond the Rule of Five: If we sample 8 items, there is a 99.2% chance that the median of the population falls within the largest and smallest values. If we take the *2nd* largest and smallest values (out of 8 total values), we get something close to a 90% CI for the median. Hubbard generalizes the tool with this handy reference table:

And if the distribution is symmetrical, then the mathless table gives us a 90% CI for the mean as well as for the median.

Catch-recatch

How does a biologist measure the number of fish in a lake? SHe catches and tags a sample of fish – say, 1000 of them – and then releases them. After the fish have had time to spread amongst the rest of the population, she’ll catch another sample of fish. Suppose she caught 1000 fish again, and 50 of them were tagged. This would mean 5% of the fish were tagged, and thus that were about 20,000 fish in the entire lake. (See Hubbard’s book for the details on how to calculate the 90% CI.)

Spot sampling

The fish example was a special case of a common problem: population proportion sampling. Often, we want to know what proportion of a population has a particular trait. How many registered voters in California are Democrats? What percentage of your customers prefer a new product design over the old one?

Hubbard's book discusses how to solve the general problem, but for now let's just consider another special case: spot sampling.

In spot sampling, you take random snapshots of things rather than tracking them constantly. What proportion of their work hours do employees spend on Facebook? To answer this, you "randomly sample people through the day to see what they were doing *at that moment*. If you find that in 12 instances out of 100 random samples" employees were on Facebook, you can guess they spend about 12% of their time on Facebook (the 90% CI is 8% to 18%).

Clustered sampling

Hubbard writes:

> "Clustered sampling" is defined as taking a random sample of groups, then conducting a census or a more concentrated sampling within the group. For example, if you want to see what share of households has satellite dishes… it might be cost effective to randomly choose several city blocks, then conduct a complete census of everything in a block. (Zigzagging across town to individually selected households would be time consuming.) In such cases, we can't really consider the number of [households] in the groups… to be the number of random samples. Within a block, households may be very similar… [and therefore] it might be necessary to treat the effective number of random samples as the number of blocks…

Measure to the threshold

For many decisions, one decision is required if a value is above some threshold, and another decision is required if that value is below the threshold. For such decisions, you don't care as much about a measurement that reduces uncertainty in general as you do about a measurement that tells you which decision to make based on the threshold. Hubbard gives an example:

> Suppose you needed to measure the average amount of time spent by employees in meetings that could be conducted remotely… If a meeting is among staff members who communicate regularly and for a relatively routine topic, but someone has to travel to make the meeting, you probably can conduct it remotely. You start out with your calibrated estimate that the median employee spends between 3% to 15% traveling to meetings that could be conducted remotely. You determine that if this percentage is actually over 7%, you should make a significant investment in tele meetings. The [EVPI] calculation shows that it is worth no more than $15,000 to study this. According to our rule of thumb for measurement costs, we might try to spend about $1,500…

> Let’s say you sampled 10 employees and… you find that only 1 spends less time in these activities than the 7% threshold. Given this information, what is the chance that the median time spent in such activities is actually below 7%, in which case the investment would not be justified? One “common sense” answer is 1/10, or 10%. Actually… the real chance is much smaller.

Hubbard shows how to derive the real chance in his book. The key point is that “the uncertainty about the threshold can fall much faster than the uncertainty about the quantity in general.”

Regression modeling

What if you want to figure out the cause of something that has many possible causes? One method is to perform a *controlled experiment*, and compare the outcomes of a test group to a control group. Hubbard discusses this in his book (and yes, he’s a Bayesian, and a skeptic of p-value hypothesis testing). For this summary, I’ll instead mention another method for isolating causes: regression modeling. Hubbard explains:

> If we use regression modeling with historical data, we may not need to conduct a controlled experiment. Perhaps, for example, it is difficult to tie an IT project to an increase in sales, but we might have lots of data about how something *else* affects sales, such as faster time to market of new products. If we know that faster time to market is possible by automating certain tasks, that this IT investment eliminates certain tasks, and those tasks are on the critical path in the time-to-market, we can make the connection.

Hubbard’s book explains the basics of linear regressions, and of course gives the caveat that correlation does not imply causation. But, he writes, “you should conclude that one thing causes another only if you have some *other* good reason besides the correlation itself to suspect a cause-and-effect relationship.”

Bayes

Hubbard’s 10th chapter opens with a tutorial on Bayes’ Theorem. For an online tutorial, see here.

Hubbard then zooms out to a big-picture view of measurement, and recommends the “instinctive Bayesian approach”:

1. Start with your calibrated estimate.
2. Gather additional information (polling, reading other studies, etc.)
3. Update your calibrated estimate subjectively, without doing any additional math.

Hubbard says a few things in support of this approach. First, he points to some studies (e.g. El-Gamal & Grether (1995)) showing that people often reason in roughly-Bayesian ways. Next, he says that in his experience, people become better intuitive Bayesians when they (1) are made aware of the base rate fallacy, and when they (2) are better calibrated.

Hubbard says that once these conditions are met,

> [then] humans seem to be mostly logical when incorporating new information into their estimates along with the old information. This fact is extremely useful because a human can consider qualitative information that does not fit in standard statistics. For example, if you were giving a forecast for how a new policy might change "public image" – measured in part by a reduction in customer complaints, increased revenue, and the like – a calibrated expert should be able to update current knowledge with "qualitative" information about how the policy worked for other companies, feedback from focus groups, and similar details. Even with sampling information, the calibrated estimator – who has a Bayesian instinct – can consider qualitative information on samples that most textbooks don't cover.

He also offers a chart showing how a pure Bayesian estimator compares to other estimators:

Also, Bayes' Theorem allows us to perform a "Bayesian inversion":

> Given a particular observation, it may seem more obvious to frame a measurement by asking the question "What can I conclude from this observation?" or, in probabilistic terms, "What is the probability X is true, given my observation?" But Bayes showed us that we could, instead, start with the question, "What is the probability of this observation if X were true?"
>
> The second form of the question is useful because the answer is often more straightforward and it leads to the answer to the other question. It also forces us to think about the likelihood of different observations given a particular hypothesis and what that means for interpreting an observation.
>
> [For example] if, hypothetically, we know that only 20% of the population will continue to shop at our store, then we can determine the chance [that] exactly 15 out of 20 would say so… [The details are explained in the book.] Then we can invert the problem with Bayes' theorem to compute the chance that only 20% of the population will continue to shop there given [that] 15 out of 20 said so in a random sample. We would find that chance to be very nearly zero…

Other methods

Other chapters discuss other measurement methods, for example prediction markets, Rasch models, methods for measuring preferences and happiness, methods for improving the subjective judgments of experts, and many others.

Step 5: Make a decision and act on it

The last step will make more sense if we first "bring the pieces together." Hubbard now organizes his consulting work with a firm into 3 phases, so let's review what we've learned in the context of his 3 phases.

Phase 0: Project Preparation

- *Initial research*: Interviews and secondary research to get familiar on the nature of the decision problem.
- *Expert identification*: Usually 4–5 experts who provide estimates.

Phase 1: Decision Modeling

- *Decision problem definition*: Experts define the problem they're trying to analyze.
- *Decision model detail*: Using an Excel spreadsheet, the AIE analyst elicits from the experts all the factors that matter for the decision being analyzed: costs and benefits, ROI, etc.
- *Initial calibrated estimates*: First, the experts undergo calibration training. Then, they fill in the values (as 90% CIs or other probability distributions) for the variables in the decision model.

Phase 2: Optimal measurements

- *Value of information analysis*: Using Excel macros, the AIE analyst runs a value of information analysis on every variable in the model.
- *Preliminary measurement method designs*: Focusing on the few variables with highest information value, the AIE analyst chooses measurement methods that should reduce uncertainty.
- *Measurement methods*: Decomposition, random sampling, Bayesian inversion, controlled experiments, and other methods are used (as appropriate) to reduce the uncertainty of the high-VoI variables.
- *Updated decision model*: The AIE analyst updates the decision model based on the results of the measurements.
- *Final value of information analysis*: The AIE analyst runs a VoI analysis on each variable again. As long as this analysis shows information value much greater than the cost of measurement for some variables, measurement and VoI analysis continues in multiple iterations. Usually, though, only one or two iterations are needed before the VoI analysis shows that no further measurements are justified.

Phase 3: Decision optimization and the final recommendation

- *Completed risk/return analysis*: A final MC simulation shows the likelihood of possible outcomes.
- *Identified metrics procedures*: Procedures are put in place to measure some variables (e.g. about project progress or external factors) continually.
- *Decision optimization*: The final business decision recommendation is made (this is rarely a simple "yes/no" answer).

Final thoughts

Hubbard’s book includes two case studies in which Hubbard describes how he led two fairly different clients (the EPA and U.S. Marine Corps) through each phase of the AIE process. Then, he closes the book with the following summary:

- If it’s really that important, it’s something you can define. If it’s something you think exists at all, it’s something you’ve already observed somehow.
- If it’s something important and something uncertain, you have a cost of being wrong and a chance of being wrong.
- You can quantify your current uncertainty with calibrated estimates.
- You can compute the value of additional information by knowing the “threshold” of the measurement where it begins to make a difference compared to your existing uncertainty.
- Once you know what it’s worth to measure something, you can put the measurement effort in context and decide on the effort it should take.
- Knowing just a few methods for random sampling, controlled experiments, or even merely improving on the judgments of experts can lead to a significant reduction in uncertainty.

How I Am Productive

I like to think that I get a lot of stuff done. Other people have noticed this and asked me how I'm so productive. This essay is where I try and "share my secrets", so to speak.

The real secret is that, in the past, I wasn't nearly as productive. I struggled with procrastination, had issues completing assignments on time, and always felt like I never had enough time to do things. But, starting in January and continuing for the past eight months, I have slowly implemented several systems and habits in my life that, taken together, have made me productive. Productivity is not a talent I have -- I've *learned* to be productive over the past several months and I have habits in place where I basically *cannot fail* to be productive.

Hopefully these systems will work for you. I've seen some people adopt them to some success, but I've never seen anyone do it *exactly* the way I do. And perhaps it would even be bad to do it exactly the way I do, because everyone is just a little bit different. I'm being aware of other-optimizing and letting you just know what's worked for me. I make no claims that these systems will work for you. Your mileage may vary.

So what are the systems? To get you to be productive, we'll need to get you to **organize**, to **prioritize**, then to **do** and **review**. Have those four things down and you'll have everything you need to be productive.

Organize

The first step to being productive **is to be organized and remember things without memorizing them.** If we get these systems down, you won't forget your ideas, when and where events are, what tasks you need to complete, what papers you have, and what emails you have.

The Most Important Rule: Write Things Down

If you only take away one system from one category, I want it to be this one. Whole essays can be written about these systems and this one is no different -- **write things down.** Whenever you have a cool idea, an event invitation, a task, etc., write it down. Always. Constantly. No excuses.

I've found in my life that stress has come in surprising part from trying to keep everything in my head. When I write down everything I think is worth remembering, whether it be a concrete thing I need to do or just a cool yet unimportant idea I want to follow up on sometime later, I write it down. That gets it out of my head, and I no longer feel the need to remember things (as long as I remember to look them up later), and I feel much better.

I've also found in my life that I constantly think I'll remember something and it's not worth writing down. More than half the time, I've been wrong and forgotten the thing. This has meant I've forgotten cool ideas and even forgotten events or to complete key items. Always write things down, no matter how convinced you are that you'll remember them.

How do you do this? I suggest getting something that will always be with you that you can write things down on. For the vast majority of my readers, this can be a phone where you text yourself messages. For a long time, I would use my smartphone to email myself notes, because I knew I'd always check my email later and then could record the note to a text document. Later on, I moved to keeping track of ideas on Evernote and then later moved on to keeping track of ideas on Workflowy. Workflowy costs $5 a month to use it to full potential (worth it, in my opinion), but there are free alternatives (that aren't as good, in my opinion).

However, don't shy away from the good old pen and paper if it gets the job done. I got this notepad for $6 and it's been great.

Keep Track of Events: The Calendar

Of course, some of the things you want to write down will be particular things that need to be recorded in particularly useful places. One of these things is **events**, or places you need to be at a particular time and place. For this, you can use any calendar, but I like Google Calendar the best. Whenever you get invited to an event, record it on your calendar. (We'll include reviewing your calendar regularly in a bit, so you won't forget what's there.)

A common mistake I see people make is to rely on Facebook events to keep track of their events. Perhaps this works for some people, but not all events are done through Facebook or can be done through Facebook, so you end up keeping track of events in multiple places, which causes confusion and missed events. Wherever you record events, **record all your events in one place.**

Keep Track of Tasks: The To-Do List

The next thing you'll want to keep track of is **tasks**. For this, you need a to-do list. I spent a lot of my life just using a TextEdit document, but I recommend you use a dedicated app instead. I personally use Workflowy here too, but others work great. In the past I've used Trello to great success. I've seen others succeed with Asana or even just a text document on the computer.

A common mistake I see people make here is using their email as their to-do list. This might make some sense, but often emails contain information unnecessary to your tasks which slows you down, and sometimes emails contain multiple action points. Worse, emails contain no easy way to prioritize tasks (which is really important and will be discussed in a bit).

Bottom line: **Keep all your tasks in one crisp, clear place.** Don't spread out your to-do lists across multiple applications and don't put it in with a bunch of other stuff.

Action, Waiting, Reference: Stay Organized with Zones

Once you have your ideas written down, your events on your calendar, and your tasks on your to-do list, it's time to organize the materials you'll have to deal with. Lots of physical papers and computer documents come at you throughout your day and it's time to organize them.

The trick here? Get a surface area you can keep clear and divide it into three zones: **action**, **waiting**, and **reference**.

The **action zone** is for things that need to be done. Have a form you need to fill out? Something you need to read? Even more outlandish things like a necklace you need to repair or something? Keep everything needed for a task together in folders or with paperclips as necessary, put it in the action zone, and record the task on your to-do list.

The **waiting zone** is for things that eventually need to be done, but which cannot be done yet because you're waiting on something. Perhaps you need feedback from someone, a package still needs to arrive, or the task only can be done on a certain day. For this, keep everything grouped together in the waiting zone, and record on your to-do list what the task is and what you're waiting for. (We'll revisit implementing zones in the to-do list in a little bit.) Move things to action and update your to-do list when what you're waiting for arrives.

The **reference zone** is for things you might need to look at and need to be kept around, but are not associated with any task. For examples, things I have had in my reference zone are passwords, details about tasks from people, items that are relevant but not necessary to the work that I'm doing, etc.

Always Inbox Zero: Apply the Folders to Your Email

Email is really messy for most people, but it doesn't have to be. The solution here is to implement the zones in your email too. I use Gmail, but nearly every email system includes folders these days. Use that system to create three folders -- action, waiting, and reference -- in your email, then sort your email according to the folders and record on your to-do list.

There is no reason to have any email in your inbox. You should be at "inbox zero" *constantly*. Whenever an email comes in, process it and file it. Got an email from Nancy that you need to reply to? Put it in "Action" and put "Reply to Nancy's email" on your to-do list. Got a long email from your boss that you don't even have time to read yet? Put it in "Action" and put "Read boss's email" on your to-do list. Then when you go back to read it, you can determine the next action item.

Emails also make sense to be put in waiting. If it's important I get a reply from the email, I'll put it in waiting to remind myself to follow up later if necessary (more on that later). I'll also put emails in waiting if I'm expecting a reply from someone else first, or if it's information for an action item I can't act on yet, or if I want to reply later on.

Lastly, reference is very important for emails that you need to keep around to read, but don't need to reply to. Lots of notes that people send me get processed into my relevant Workflowy document and then kept in reference for as long as they're relevant.

Prioritize

Now that you're all organized, it's time to get in a position to do the things you need to do. But watch out, because unless you have time to complete your entire to-do list in one sitting, it's a poor use of time to just go from the top to the bottom. Instead, we need to **go from the most important to the least important.**

Eisenhower Matrix: Do What's Important

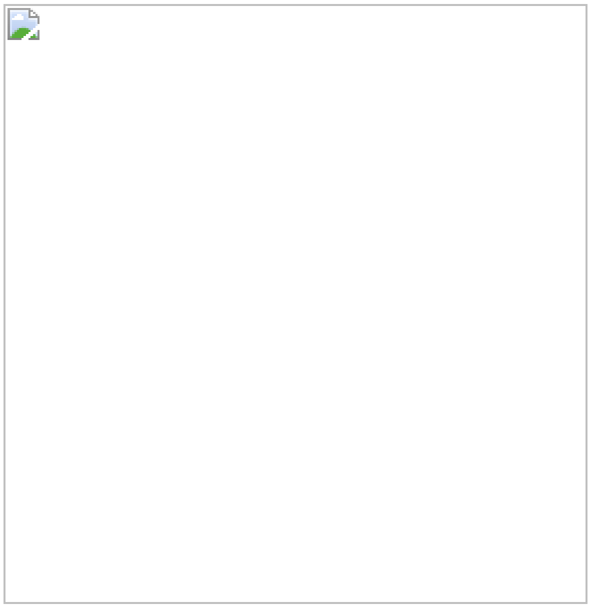
How do you prioritize? The best tactic I've seen here is called The Eisenhower Matrix. It comes from Steven Covey's book **First Things First** but is credited to President Dwight D. Eisenhower.

Here, you take your to-do list and organize everything into four quadrants: important and urgent, important and not urgent, unimportant and urgent, and unimportant and not urgent. This is very easy to do on Workflowy, and still possible on something like Trello.

There's pretty universal agreement that you complete all the "important and urgent" tasks first and the "unimportant and not urgent" tasks last. But the real trick is that **after you complete the important and urgent tasks, you should move to complete the important and not urgent tasks.** Ignore the not important and urgent tasks until you've completed all important tasks and even be comfortable with skipping unimportant tasks if necessary. Why? Because they're not important.

If you get this matrix down, you'll soon get ahead on your tasks, because you'll be completing important tasks *before they become urgent*.

Also, note the inclusion of "waiting" here as one of the tabs in my to-do list. This is where I put tasks I can't complete yet with a note of what I'm waiting on. Something like talking to my Dad three days from now would be tagged as "#30aug :: Talk to Dad" (using Workflowy hashtags), but I'd also do things with unclear dates, like "Brian responds to email :: Forward response to Seth". Beware that being able to manage unclear deadlines (where you don't know what day the task will be) is something that most to-do list apps struggle with.



Timeboxing: Plan Your Day in Advance

The next prioritization thing to master is *planning your day in advance*. You do this through making "time boxes" for things, or periods of time where you'll do something predefined. For example, I'll set aside some time to work through my to-do list or to work on particular projects. For bigger projects, I'll decide how much I want to work on them in any particular day or week and set them aside from my to-do list. I'll then block out time for them on my calendar and end up with days like this.

Since I plan my days in advance using this timebox method, I just plan every minute of the calendar in advance and have a plan so I always know what to be doing and never miss a beat. Of course, things come up and you'll have to change your plan for the day, but that's better than having no plan at all.

Two Minute Rule

It's important to be mindful of how much time it takes to record a task, put it in your to-do list, and prioritize it, however. For most people, including me, it's about two minutes for any given task. This gives rise to the "two minute rule": **if doing somethign would take less than two minutes, just do it now.** Likewise, if it would take over two minutes, put it in your to-do list and do it at the best time.

Do

Now that you have your to-do list set and timeboxes for when you're going to work and on what, it's time to actually *do the work*.

The Pomodoro Technique

The ideal timebox should be a length that is a multiple of thirty minutes so you can do the most powerful productivity thing there is: The Pomodoro Technique. Beware that it doesn't work for some, but I do urge you to give it a fair shake and a few tries, because for those whom the Pomodoro works, the Pomodoro Technique works wonders.

Here's how you do it. Set a timer for 25 minutes. During those 25 minutes (a) work **only** on your task at hand; (b) **do not** do anything else, even for a second; (c) be completely focused; (d) be free from distractions; (e) and **do not multi-task.** There are some acceptable things to do during a Pomodoro, however: go to the bathroom, drink, listen to music. But there are tons more things not to do during a Pomodoro: check Facebook, read your email, etc. The list will go on.

After the timer expires, take a five minute break. During these five minutes, do anything you'd like **except** the task on hand. Even if you feel like the break is boring and you're itching to get back on task, **don't.** You're only hurting yourself in the long-run. This five minute break will restore your focus, keep you grounded, provide a way to think through your ideas in a different setting, and prevent you from needing longer breaks later in the day.

It should be noted, however, that the Pomodoro can be a bit difficult to get in the habit of, though. To solve this, I've found it useful to work my way up to the full Pomodoro by spending a month getting used to "15 minutes of work, 5 minutes break", then another month doing "20 minutes of work, 5 minutes break", and then finally "25 minutes of work, 5 minutes break".

Different people have tried other multiples besides 25 and 5, but I'm still convinced that 25-5 is the ideal split. Perhaps 27-3 could work better for advanced Pomodoro users, but I wouldn't push it further. I've seen things like 90-30 or 30-10, and all of these seem to involve working just a little too long (losing focus) and then taking a lot more break than is necessary. Of course, if it works for you, then it works.

Here's the 25-5 stopwatch I use and my 20-5 stopwatch. I've also liked Tomato Timer.com, but any timer can work.

Be Comfortable with Breaks

The important lesson of working a lot is to be comfortable with taking a break. The novice productive person will think it virtuous to work clear through a break and onward, thinking that he or she is making even better use of their time, defeating all those sissy workers who need breaks! But really, this person is just setting up their own downfall, because they'll crash and burn.

Burnout is real and one of the most dangerous things you can do is train yourself to feel guilty about not working. So you need to remember to take breaks. The break in a Pomodoro is a good one, but I also recommend taking a larger break (like 30 minutes) after completing three or four Pomodoros.

One particularly good break I'd like to give a shout-out to is to take a nap. Taking a nap at a fairly regular time has health benefits (see also here, here, and here) and doesn't harm your night sleep if you nap for 20 minutes and don't nap too late in the afternoon or evening. In fact, I've actually found naps to be a *time saver* instead of time "wasted" for a break, because I can sleep less at night and still feel rested and be focused throughout the day.

Keep Your Energy Up

Another thing to prevent your chance of crashing and needing a long break to restore your energy is to keep your energy up. I recommend drinking something that is somewhat sugary but not too sugary (I drink water-diluted lemonade in a 25%-75% mix) and remembering to exercise on a regular basis. Also, eating healthy and sleeping right works wonders for keeping your attention on your work.

Review

Of course, it's not enough to do if you're not going to learn from how you're doing and improve. I suggest you review your life on multiple levels -- daily, weekly, monthly, and once every six months.

For the **daily review**, I keep track of whether I've succeeded at certain habits like exercising and eating right, and log the amount of time I've spent on various things so I can keep track of my time usage. I also complete other relevant logs, and then spend a bit of time reflecting how things have gone for the day and think of ways to repeat successes and avoid mistakes. I then check the plan for the next day and tweak it if necessary. This process takes me about 15 to 20 minutes.

For the **weekly review**, I go through my action-waiting-reference zones wherever they exist (physical piles, email, and computer folders) and process them -- make sure everything there is still relevant and still belongs in the same place. I'll remove whatever needs to be removed at this stage and remind myself what I'm working on. I'll organize and clean anything that isn't organized at this stage and get everything together. I'll then quickly re-read my strategic plan and plan out the week in accordance with my goals. Recently, I've set amounts of time per week I want to be spending on certain projects, so it's now a matter of making a schedule that works. This process usually takes me 45 minutes to an hour.

For the **monthly review**, I reflect on the habits I've been trying to build for the month and decide what habits I want to keep, what habits I want to add, and what habits I want to subtract. I review how the month as a whole went and think about what I can do to repeat successes and avert future failures. I then write up a reflection and publish it on my blog. This process usually takes me two hours.

For the **six month review**, I return to my goals and think about how my life trajectory as a whole is going. What are my life goals? What am I doing to accomplish them? Am I closer to my goals than I was six months ago? Should I be working toward new goals? What common mistakes did I make through the past six months that I want to avoid? I then write a documentwith my personal mission and goals for the next six months and skim it every week to constantly remind myself of what I want to be doing. This process usually takes me three hours.

Yes, there will be an unlucky day where you do all four reviews and spend like six and a half hours reviewing your life at different levels. Perhaps this is a bit much for people, but I've found tremendous benefit from it. I've found that spending this day reviewing my life has saved me from not just days, but even months, of wasted time that doesn't accomplish what I really want to do. **Reviewing is another way of saving you time.**

Additional Tips

Now I've given you all my main advice, but I have some additional tips if you want to keep reading.

Carefully form these habits over time. This is a lot to do at once, so do it in stages. Build the habit of writing things down first, and then slowly get the apps you like in place for ideas, events, and tasks. After you have that down, spend the time necessary to get your email in order and implement the zones wherever possible. Then begin to move into prioritizing your tasks with the Eisenhower Matrix. After you have this down, begin planning your days in advance with timeboxes and start doing your reviews. While you're building that habit, simultaneously start building up the Pomodoro habit, slowly approaching 25-5 over a few months.

Find a way to reliably stay on habit. Don't make the common failure of sticking to something for a month or two and abandoning it. Spend a lot of energy thinking through how you'll stay on habit and how you'll not be like all the other people who think they'll stay on habit then fail. Make a bet with a friend, start up Beeminder, or create some other kind of commitment device.

Form the productivity mindset. I had a lot of trouble implementing this plan until I was able to think of myself as an important person who does important things and should personally value my time. I had to really want to be productive before I could start being productive. Success at this will follow from the right mindset. It's time to start thinking of yourself as important. If you can't fool yourself, maybe it's time to look at your goals and decide what goals would make you feel important and then *do those goals instead*.

Behold the power of routines. I find it a lot easier to exercise if I have a routine of "every other day, right after waking up" or "every other day, right before dinner".

Your routine can be built from here. It's a lot easier to stick to timeboxes if they're regularly occurring. Use a calendar and build yourself something nice.

Put everything in a particular place. People lose a lot of time just hunting around for things. Solve this by spending some time ahead of time organizing things in your life and getting them into particular places. Then always make sure things return to their places.

Declutter your life. You'll work better if you have less stuff to keep track of and less commitments to worry about. Get rid of everything and delegate anything you can.

Make a productivity place. This works especially well in colleges where there is a large variety of places you could be working. Find a place to work, set up your Pomodoros, and follow them to the letter. Don't mess up. Take your longer breaks somewhere else. If you do mess up, find a new productivity place and start again. I found this really helpful for my mindset, but others have found it silly.

Don't neglect friends and family. This is a big one. Remember, the goal of being more productive is to free time to do the things you want and be with the people you want. It's not to spend 100 hour workweeks neglecting those who are important to you. Make sure to take some time off to spend with friends and family. Schedule it in your calendar if you have to. This will matter most in the long-run for your life.

Productivity ≠ Busy and Busy ≠ Productivity. If you do productivity right, you shouldn't feel busy all that often. Being busy is a sign of having poor productivity and/or having taken on too many commitments, and is rarely ever a sign of doing things correctly.

Conclusion

These tips are really a result of me experimenting for eight months. I'd expect you to take a similar amount of time to go from zero to productive and end up with different systems that work for you and your environment. But I think there are a lot of power in these systems and I'm interested to see what other people do and how other people run with them. After all, they work for *me*.

Further reading:

* The Secret Weapon

* David Allen's **Getting Things Done: The Art of Stress-Free Productivity**

* Scott Young's **The Little Book of Productivity**

* Paul Christiano's Workflow

* 10 Step Anti-Procrostination Checklist

* Zenhabits

-

(Also cross-posted on my blog.)

Where I've Changed My Mind on My Approach to Speculative Causes

Follow up to Why I'm Skeptical About Unproven Causes (And You Should Be Too)

Previously, I wrote "Why I'm Skeptical About Unproven Causes (And You Should Be Too)" and a follow up essay "What Would It Take to Prove a Speculative Cause?". Both of these sparked a lot of discussion on LessWrong, on the Effective Altruist blog, and my own blog, as well as many hours of in person conversation.

After all this extended conversation with people, I've changed my mind on a few things that I will elaborate here. I hope in doing so I can (1) clarify my original position and (2) explain where I now stand in light of all the debate so people can engage with my *current ideas* as opposed to the ideas I no longer hold. My opinions on things tend to change quickly, so I think updates like this will help.

My Argument, As It Currently Stands

If I were to communicate one main point of my essay, based on what I believe now, it would be **when you're in a position of high uncertainty, the best response is to use a strategy of exploration rather than a strategy of exploitation.**

What I mean by this is that given the high uncertainty of impact we see now, especially with regard to the far future, we're better off trying to find more information about impact and reduce our uncertainty (exploration) rather than pursuing whatever we think is best (exploitation).

The implications of this would mean that:

- We should develop more of an attitude that our case for impact is neither clear nor proven.
- We should apply more skepticism to our causes and more self-skepticism to our personal beliefs about impact.
- We should use the language of "information" and "exploration" more often than the language of "impact" and "exploitation".
- We should focus more on finding specific and concrete attempts to ensure we're making progress and figure out our impact (whether it be surveys, experiments, soliciting external review from relevant experts, etc.).
- We should focus more on transparency about what we're doing and thinking and why, when relevant and not exceedingly costly.

And to be clear, here are specific statements that address misconceptions about what I have argued:

- I do think it is wrong to ignore unproven causes completely and stop pursuing them.
- I don't think we should be donating everything to the Against Malaria Foundation instead of speculative causes.
- I don't think the Against Malaria Foundation has the highest impact of all current opportunities to donate.
- I do think we can say useful things about the far future.
- I don't think the correct way to think about high uncertainty and low evidence is to "suspend judgement". Rather, I think we should make a judgement that we expect the estimate to be much lower than initially claimed in light of all the things I've said earlier about the history of past cost-effectiveness estimates.

And, lastly, if I were to make a second important point it would be **it's difficult to find good opportunities to buy information.** It's easy to think that any donation to an organization will generate good information or that we'll automatically make progress just by working. I think some element of random pursuit is important (see below), but all things considered I think we're doing too much random pursuit right now.

Specific Things I Changed My Mind About

Here are the specific places I changed my mind on:

I used to think donating to AMF, at least in part, was important for me. Now I don't.

I underestimated the power of exploring and the existing opportunities, so I think that 100% of my donations should be going to trying to assess impact. I've been persuaded that there is already quite a lot of money going toward AMF and we might not need more money as quickly as thought, so for the time being it's probably more appropriate to save and then donate to opportunities to buy information as they come up.

I now agree that there are relevant economies of scale in pursuing information that I hadn't taken into account.

What I mean by this is it might not be appropriate for individuals to work on purchasing information themselves. Instead, this could end up splitting up the time of organizations unnecessarily as they provide information to a bunch of different people. Also, many people don't have the time to do this themselves.

I think this has two implications:

- We should put more trust in larger scale organizations who are doing exploring, like GiveWell, and pool our resources.
- Individuals should work harder to put relevant information about information we gather online.

I was partially mistaken in thinking about how to "prove" speculative causes.

I think there was some value in my essay "What Would It Take to Prove a Speculative Cause?" because it talked concretely about strategies some organizations could take to get more information about their impact.

But the overall concept is mistaken -- there is no arbitrary threshold of evidence at which a speculative cause needs to cross and I was wasting my time by trying to come up with one. Instead, I think it's appropriate to continue doing expected value calculations as long as we maintain a self-skeptical, pro-measurement mindset.

I had previously not fully taken into account the cost of acquiring further information.

The important question in value of information is not "what does this information get me in terms of changing my beliefs and actions?" but actually "how valuable is this information?", as in, do the benefits of gathering this information outweigh all the costs? In some cases, I think the benefits of further proving a cause probably don't outweigh the costs.

For one possibly extreme example, while I don't know the rationale for doing a 23rd randomized controlled trial on anti-malaria bednets after performing the previous 22, it's likely that doing that RCT would have to be testing something more specific than the general effectiveness of bednets to justify the high cost of doing an RCT.

Likewise, there are costs on organizations to devoting resources to measuring themselves and being more transparent. I don't think these costs are particularly high or defeat the idea of devoting more resources to this area, but I hadn't really taken them into account before.

I'm slightly more in favor of acting randomly (trial and error).

I still think it's difficult to acquire good value of information and it's very easy to get caught "spinning our wheels" in research, especially when that research has no clear feedback loops. One example, perhaps somewhat controversial, would be to point to the multi-century lack of progress on some problems in philosophy (think meta-ethics) as an example of what can happen to a field when there aren't good feedback loops to ground yourself.

However, I underestimated the amount of information that comes forward just doing ones normal activities. The implication here is that it's more worthwhile than I initially thought to fund speculative causes just to have them continue to scale and operate.

-

(This was also cross-posted on my blog.)

Reality is weirdly normal

Related to: When Anthropomorphism Became Stupid, Reductionism, How to Convince Me That 2 + 2 = 3

"**Reality is normal.**" That is: Surprise, confusion, and mystery are features of maps, not of territories. If you would think like reality, cultivate outrage at yourself for failing to intuit the data, not resentment at the data for being counter-intuitive.

"**Not one unusual thing has ever happened.**" That is: Ours is a tight-knit and monochrome country. The cosmos is simple, tidy, lawful. "[T]here is no surprise from a causal viewpoint — no disruption of the physical order of the universe."

"**It all adds up to normality.**" That is: Whatever is true of fundamental reality does not exist in a separate universe from our everyday activities. It *composes* those activities. The perfected description of our universe must in principle allow us to reproduce the appearances we started with.

These maxims are remedies to magical mereology, anthropocentrism, and all manner of philosophical panic. But reading too much (or too little) into them can lead seekers from the Path. For instance, they may be wrongly taken to mean that the world is obliged to validate our initial impressions or our untrained intuitions. As a further corrective, I suggest: **Reality is *weirdly* normal**. It's "normal" in odd ways, by strange means, in surprising senses.

At the risk of vivisecting poetry, and maybe of stating the obvious, I'll point out that the maxims mean different things by "normal". In the first two, what's "normal" or "usual" is the universe taken on its own terms — the cosmos as it sees itself, or as an ideally calibrated demon would see it. In the third maxim, what's "normal" is the universe *humanity* perceives — though this still doesn't identify normality with what's *believed* or *expected*. Actually, it will take some philosophical work to articulate just what Egan's "normality" should amount to. I'll start with Copernicanism and reductionism, and then I'll revisit that question.

Everything is usual. Very nearly nothing is familiar.

- Spell 1: *Since the beginning, not one unusual thing has ever happened. And intelligence explosion sounds... well, 'unusual' is the mildest word that comes to mind. So it won't happen.*
- Counterspell: Reality is *weirdly* normal. It's not the kind of normal that *sounds* normal.

Relative to the universe's laws, everything is par for the course. But relative to *human* standards of normality — what I'll call the **familiar** — *barely anything* that actually exists is par for the course. Nearly all events are senseless, bizarre, inhuman. But that's because our mapping hardware is adapted to a very specific environment; if anything's objectively weird, it's we, not the other denizens of Everythingland. That's much of why absurdity is a poor guide to probability.

Egan's Law reminds us that beneath our human surface weirdness lies a deeper regularity, a deeper unity with the rest of Nature. But to call this alien order 'regular' already assumes a shift in perspective from what a person off the street would initially think of as 'regular', to the hidden patterns of the very large and very small. We should prize the ability to shift between these points of view, while carefully avoiding conflating them.

- Spell 2: *Reality is normal. So it shouldn't be difficult for me to think like it.*
- Counterspell: What's normal from a God's-eye view is wont to be weird from a human's. And vice versa.

Thinking like reality requires *understanding* reality. You can't just will yourself into becoming a high-fidelity map. If nothing else, you'll be too vulnerable to knowledge gaps. *Pretending* you think like quantum physics is even worse than rebelling against the physics for being too confusing. At least in the latter case you've noticed the disparity between your map and the territory.

To take pride in one's confusion at the quantum, rather than striving mightily to understand, is an epistemic sin. But to *deny* one's confusion at the quantum, to push it from one's mind and play-act at wisdom, is a graver sin. The goal isn't to do away with one's confusion; it's to do away with one's best *reasons* to be confused.

What adds up to the familiar needn't be familiar.

- Spell 3: *It all adds up to normality. But relativity sure sounds abnormal! So relativity will be replaced by something normaler.*
- Counterspell: Physics is the weird addenda, not the normal sum.

x + y + z + ... = Normality. But x ≠ Normality. Nor does y. In fact, even the complete quantum-mechanical description of the human world would not *look* normal, from a human perspective. But it would be a description of *exactly* the world we live in.

This is another reminder that surprise is a feature of maps, not of territories. Very different maps — a wave function scribbled in silicon, a belief, a gesture, a child's drawing — can represent the same territory. The drawing is normal (familiar), and the wave function isn't. But the *referent* of the two maps may well be the same. In asserting "quantum mechanics adds up to normality", we're really asserting that our maps of familiar objects, to the extent they're accurate, would co-refer with portions of the maps of an ideal Finished Science. They're two different languages, but faithful translation is possible.

- Spell 4: *It all adds up to normality. So I should be able to readily intuit how QM yields the familiar.*
- Counterspell: Reality forms what's normal, but by weird means.

The *relationship* between the quantum and the familiar can be thoroughly unfamiliar. The way in which the unintuitive fundamental truths yield an intuitive Middle World is not itself intuitive.

We could call this "emergence", if we wished to craft a territory predicate out of mapstuff. Otherwise, I suggest calling this predicate"wow-sometimes-I'm-not-very-good-at-keeping-track-of-lots-of-small-things-ence". Understanding the logical or mathematical implications of simple patterns is not humanity's specialty. Causal and part-whole relations are no exception to this rule.

It all adds up to *the phenomenon*.

- Spell 5: *It all adds up to normality. It's common sense that time 'flows'. So science will ultimately show time is not at all like space.*
- Counterspell: We'd find even the familiar alien, if we but understood it.

Normality isn't common sense. Normality isn't our beliefs, expectations, assumptions, or strongest convictions. Normality, in the sense relevant to a *true and binding* "It all adds up to normality", is the phenomenon, the human world as it appears. (Not the human world as it is *believed* to appear. The human world as we *actually* encounter it.)

What's "the phenomenon"?

I'm going to be willfully cagey about that. What I really mean by "the phenomenon" is "whatever's going on". Or whatever's going on that we're cognitively accessing or representing. It needn't all add up to a world that makes my map true; but it *will* add up to a world that *explains why my map looks the way it does*.

Even saying that much risks overly restricting the shape explanations are allowed to take. Egan's Law can be used as a general constraint on successful explanations — they must account for the *explanandum* — but only if we treat the *explanandum* loosely enough to permit dissolutions and eliminations alongside run-of-the-mill reductions. It all adds up to an explanation of how our beliefs arose, though not necessarily a validation of them.

- Spell 6: *It all adds up to normality. I have an immediate epistemic acquaintance with the irreducible phenomenal character of my experience. So, whatever be the final theory, it will certainly include qualia.*
- Counterspell: Every map is in a territory. But no meta-map is beyond suspicion.

It's obvious that appearances can deceive us about underlying reality. But a variety of perceptual illusions demonstrate that experiences can also consistently mislead us about *themselves*. It all adds up to normality, but normality *lies*.

In some cases, we can't accurately describe our *surface impressions* until we've understood their underlying mechanism. Since so much of science is revisionary, we mustn't interpret Egan's Law in a way that unduly privileges first-pass descriptions. When data and theory conflict, it's sometimes more likely that the data has been misrepresented than that the theory is false.

How far does this openness to map revision extend? As far as the dynamics of one's brain allows. From the sidelines, it may appear to me that *in principle* a thinker should be able to have certainty about some propositions — for instance, 'an experience is occurring'. But when I actually find myself living through such a thought, I don't in fact *experience* infinite confidence. It remains physically possible for me to be persuaded otherwise.

This point ought to be a bit controversial, and more defense of it is needed. But I do insist on treating 'phenomenal experience' and 'phenomenon' as *prima facie* independent concepts. Egan's Law really is *the law*, whereas claims about some inerrant mode of reasoning or perception will at best qualify as well-supported hypotheses.

Egan's Law is not about saving conscious experience, or our theories, or our axioms, or our interpretations or descriptions of the phenomenon. It's about **saving *the phenomenon itself*** — the piece of the world in which we are, in fact, submerged. Egan's Law can be restated: The part of reality that (under its familiar description) puzzled or surprised us is identical to (or otherwise lawfully derivable from) the part of reality that (under its more fundamental description) explains it.

The human piece of the universe is of a piece with everything else. And the Everything Else gets explanatory priority. Of all our science's findings to date, that may well be the most startling.

Common sense as a prior

Introduction

[I have edited the introduction of this post for increased clarity.]

This post is my attempt to answer the question, "How should we take account of the distribution of opinion and epistemic standards in the world?" By “epistemic standards,” I roughly mean a person’s way of processing evidence to arrive at conclusions. If people were good Bayesians, their epistemic standards would correspond to their fundamental prior probability distributions. At a first pass, my answer to this questions is:

> **Main Recommendation**: Believe what you think a broad coalition of trustworthy people would believe if they were trying to have accurate views and they had access to your evidence.

The rest of the post can be seen as an attempt to spell this out more precisely and to explain, in practical terms, how to follow the recommendation. Note that there are therefore two broad ways to disagree with the post: you might disagree with the main recommendation, or the guidelines for following main recommendation.

The rough idea is to try find a group of people whose are trustworthy by clear and generally accepted indicators, and then use an impartial combination of the reasoning standards that they use when they are trying to have accurate views. I call this impartial combination *elite common sense*. I recommend using elite common sense as a prior in two senses. First, if you have no unusual information about a question, you should start with the same opinions as the broad coalition of trustworthy people would have. But their opinions are not the last word, and as you get more evidence, it can be reasonable to disagree. Second, a complete prior probability distribution specifies, for any possible set of evidence, what posterior probabilities you should have. In this deeper sense, I am not just recommending that you start with the same opinions as elite common sense, but also you update in ways that elite common sense would agree are the right ways to update. In practice, we can’t specify the prior probability distribution of elite common sense or calculate the updates, so the framework is most useful from a conceptual perspective. It might also be useful to consider the output of this framework as one model in a larger model combination.

I am aware of two relatively close intellectual relatives to my framework: what philosophers call “equal weight” or “conciliatory” views about disagreement and what people on LessWrong may know as “philosophical majoritarianism.” Equal weight views roughly hold that when two people who are expected to be roughly equally competent at answering a certain question have different subjective probability distributions over answers to that question, those people should adopt some impartial combination of their subjective probability distributions. Unlike equal weight views in philosophy, my position is meant as a set of rough practical guidelines rather than a set of exceptionless and fundamental rules. I accordingly focus on practical issues for applying the framework effectively and am open to limiting the framework’s scope of application. Philosophical majoritarianism is the idea that on most issues, the average opinion of humanity as a whole will be a better guide to the truth than one’s own personal judgment. My perspective differs from both equal weight views and philosophical majoritarianism in that it emphasizes an elite subset of the population rather than humanity as a whole and that it emphasizes epistemic standards more than individual opinions. My perspective differs from what you might call "elite majoritarianism" in that, according to me, you can disagree with what very trustworthy people think on average if you think that those people would accept your views if they had access to your evidence and were trying to have accurate opinions.

I am very grateful to Holden Karnofsky and Jonah Sinick for thought-provoking conversations on this topic which led to this post. Many of the ideas ultimately derive from Holden's thinking, but I've developed them, made them somewhat more precise and systematic, discussed additional considerations for and against adopting them, and put everything in my own words. I am also grateful to Luke Muehlhauser and Pablo Stafforini for feedback on this post.

In the rest of this post I will:

1. Outline the framework and offer guidelines for applying it effectively. I explain why I favor relying on the epistemic standards of people who are trustworthy by clear indicators that many people would accept, why I favor paying more attention to what people think than why they say they think it (on the margin), and why I favor stress-testing critical assumptions by attempting to convince a broad coalition of trustworthy people to accept them.
2. Offer some considerations in favor of using the framework.
3. Respond to the objection that common sense is often wrong, the objection that the most successful people are very unconventional, and objections of the form "elite common sense is wrong about X and can't be talked out of it."
4. Discuss some limitations of the framework and some areas where it might be further developed. I suspect it is weakest in cases where there is a large upside to disregarding elite common sense, there is little downside, and you'll find out whether your bet against conventional wisdom was right within a tolerable time limit, and cases where people are unwilling to carefully consider arguments with the goal of having accurate beliefs.

An outline of the framework and some guidelines for applying it effectively

My suggestion is to use elite common sense as a prior rather than the standards of reasoning that come most naturally to you personally. The three main steps for doing this are:

1. Try to find out what people who are trustworthy by clear indicators that many people would accept believe about the issue.
2. Identify the information and analysis you can bring to bear on the issue.
3. Try to find out what elite common sense would make of this information and analysis, and adopt a similar perspective.

On the first step, people often have an instinctive sense of what others think, though you should beware the false consensus effect. If you don't know what other opinions are out there, you can ask some friends or search the internet. In my experience, regular people often have similar opinions to very smart people on many issues, but are much worse at articulating considerations for and against their views. This may be because many people copy the opinions of the most trustworthy people.

I favor giving more weight to the opinions of people who can be shown to be trustworthy by clear indicators that many people would accept, rather than people that seem trustworthy to you personally. This guideline is intended to help avoid parochialism and increase self-skepticism. Individual people have a variety of biases and blind spots that are hard for them to recognize. Some of these biases and blind spots—like the ones studied in cognitive science—may affect almost everyone, but others are idiosyncratic—like biases and blind spots we inherit from our families, friends, business networks, schools, political groups, and religious communities. It is plausible that combining independent perspectives can help idiosyncratic errors wash out.

In order for the errors to wash out, it is important to rely on the standards of people who are trustworthy by clear indicators that many people would accept rather than the standards of people that seem trustworthy to you personally. Why? The people who seem most impressive to us personally are often people who have similar strengths and weaknesses to ourselves, and similar biases and blind spots. For example, I suspect that academics and people who specialize in using a lot of explicit reasoning have a different set of strengths and weaknesses from people who rely more on implicit reasoning, and people who rely primarily on many weak arguments have a different set of strengths and weaknesses from people who rely more on one relatively strong line of argument.

Some good indicators of general trustworthiness might include: IQ, business success, academic success, generally respected scientific or other intellectual achievements, wide acceptance as an intellectual authority by certain groups of people, or success in any area where there is intense competition and success is a function of ability to make accurate predictions and good decisions. I am less committed to any particular list of indicators than the general idea.

Of course, trustworthiness can also be domain-specific. Very often, elite common sense would recommend deferring to the opinions of experts (e.g., listening to what physicists say about physics, what biologists say about biology, and what doctors say about medicine). In other cases, elite common sense may give partial weight to what putative experts say without accepting it all (e.g. economics and psychology). In other cases, they may give less weight to what putative experts say (e.g. sociology and philosophy). Or there may be no putative experts on a question. In cases where elite common sense gives less weight to the opinions of putative experts or there are no plausible candidates for expertise, it becomes more relevant to think about what elite common sense would say about a question.

How should we assign weight to different groups of people? **Other things being equal, a larger number of people is better, more trustworthy people are better, people who are trustworthy by clearer indicators that more people would accept are better, and a set of criteria which allows you to have some grip on what the people in question think is better, but you have to make trade-offs.** If I only included, say, the 20 smartest people I had ever met as judged by me personally, that would probably be too small a number of people, the people would probably have biases and blind spots very similar to mine, and I would miss out on some of the most trustworthy people, but it would be a pretty trustworthy collection of people and I'd have some reasonable sense of what they would say about various issues. If I went with, say, the 10 most-cited people in 10 of the most intellectually credible academic disciplines, 100 of the most generally respected people in business, and the 100 heads of different states, I would have a pretty large number of people and a broad set of people who were very trustworthy by clear standards that many people would accept, but I would have a hard time knowing what they would think about various issues because I haven't interacted with them enough. How these factors can be traded-off against each other in a way that is practically most helpful probably varies substantially from person to person.

I can't give any very precise answer to the question about whose opinions should be given significant weight, even in my own case. Luckily, I think the output of this framework is usually not very sensitive to how we answer this question, partly because most people would typically defer to other, more trustworthy people. If you want a rough guideline that I think many people who read this post could apply, I would recommend focusing on, say, the opinions of the top 10% of people who got Ivy-League-equivalent educations (note that I didn't get such an education, at least as an undergrad, though I think you should give weight to my opinion; I'm just giving a rough guideline that I think works reasonably well in practice). You might give some additional weight to more accomplished people in cases where you have a grip on how they think.

I don't have a settled opinion about how to aggregate the opinions of elite common sense. I suspect that taking straight averages gives too much weight to the opinions of cranks and crackpots, so that you may want to remove some outliers or give less weight to them. For the purpose of making decisions, I think that sophisticated voting methods (such as the Condorcet method) and analogues of the parliamentary approaches outlined by Nick Bostrom and Toby Ord seem fairly promising as rough guidelines in the short run. I don't do calculations with this framework—as I said, it's mostly conceptual—so uncertainty about an aggregation procedure hasn't been a major issue for me.

On the margin, I favor paying more attention to people's opinions than their explicitly stated reasons for their opinions. Why? One reason is that I believe people can have highly adaptive opinions and patterns of reasoning without being able to articulate good defenses of those opinions and/or patterns of reasoning. (Luke Muehlhauser has discussed some related points here.) One reason is that people can adopt practices that are successful without knowing why they are successful, others who interact with them can adopt those practices, others who interact with them can adopt those practices, and so forth. I heard an extreme example of this from Spencer Greenberg, who had read it in Scientists Greater than Einstein. The story involved a folk remedy for visual impairment:

> There were folk remedies worthy of study as well. One widely used in Java on children with either night blindness or Bitot's spots consisted of dropping the juices of lightly roasted lamb's liver into the eyes of affected children. Sommer relates, "We were bemused at the appropriateness of this technique and wondered how it could possibly be effective. We, therefore, attended several treatment sessions, which were conducted exactly as the villagers had described, except for one small addition—rather than discarding the remaining organ, they fed it to the affected child. For some unknown reason this was never considered part of the therapy itself." Sommer and his associates were bemused, but now understood why the folk remedy had persisted through the centuries. Liver, being the organ where vitamin A is stored in a lamb or any other animal, is the best food to eat to obtain vitamin A. (p. 14)

Another striking example is bedtime prayer. In many Christian traditions I am aware of, it is common to pray before going to sleep. And in the tradition I was raised in, the main components of prayer were listing things you were grateful for, asking for forgiveness for all the mistakes you made that day and thinking about what you would do to avoid similar mistakes in the future, and asking God for things. Christians might say the point of this is that it is a duty to God, that repentance is a requirement for entry to heaven, or that asking God for things makes God more likely to intervene and create miracles. However, I think these activities are reasonable for different reasons: gratitude journals are great, reflecting on mistakes is a great way to learn and overcome weaknesses, and it is a good idea to get clear about what you really want out of life in the short-term and the long-term.

Another reason I have this view is that if someone has an effective but different intellectual style from you, it's possible that your biases and blind spots will prevent you from appreciating their points that have significant merit. If you partly give weight to opinions independently of how good the arguments seem to you personally, this can be less of an issue for you. Jonah Sinick described a striking reason this might happen in Many Weak Arguments and the Typical Mind:

> **We should pay more attention to people's bottom line than to their stated reasons** — If most high functioning people aren't relying heavily on any one of the arguments that they give, if a typical high functioning person responds to a query of the type "Why do you think X?" by saying "I believe X because of argument Y" we shouldn't conclude that the person believes argument Y with high probability. Rather, we should assume that argument Y is one of many arguments that they believe with low confidence, most of which they're not expressing, and we should focus on their belief in X instead of argument Y. [emphasis his]

This idea interacts in a complementary way to Luke Muehlhauser's claim that some people who are not skilled at explicit rationality may be skilled in tacit rationality, allowing them to be successful at making many types of important decisions. If we are interacting with such people, we should give significant weight to their opinions independently of their stated reasons.

A counterpoint to my claim that, on the margin, we should give more weight to others' conclusions and less to their reasoning is that some very impressive people disagree. For example, Ray Dalio is the founder of Bridgewater, which, at least as of 2011, was the world's largest hedge fund. He explicitly disagrees with my claim:

> "I stress-tested my opinions by having the smartest people I could find challenge them so I could find out where I was wrong. I never cared much about others' conclusions—only for the reasoning that led to these conclusions. That reasoning had to make sense to me. Through this process, I improved my chances of being right, and I learned a lot from a lot of great people." (p. 7 of Principles by Ray Dalio)

I suspect that getting the reasoning to make sense to him was important because it helped him to get better in touch with elite common sense, and also because reasoning is more important when dealing with very formidable people, as I suspect Dalio did and does. I also think that for the some of the highest functioning people who are most in touch with elite common sense, it may make more sense to give more weight to reasoning than conclusions.

The elite common sense framework favors testing unconventional views by seeing if you can convince a broad coalition of impressive people that your views are true. If you can do this, it is often good evidence that your views are supported by elite common sense standards. If you can't, it's often good evidence that your views can't be so supported. Obviously, these are rules of thumb and we should restrict our attention to cases where you are persuading people by rational means, in contrast with using rhetorical techniques that exploit human biases. There are also some interesting cases where, for one reason or another, people are unwilling to hear your case or think about your case rationally, and applying this guideline to these cases is tricky.

Importantly, I don't think cases where elite common sense is biased are typically an exception to this rule. In my experience, I have very little difficulty convincing people that some genuine bias, such as scope insensitivity, really is biasing their judgment. And if the bias really is critical to the disagreement, I think it will be a case where you can convince elite common sense of your position. Other cases, such as deeply entrenched religious and political views, may be more of an exception, and I will discuss the case of religious views more in a later section.

The distinction between convincing and "beating in an argument" is important for applying this principle. It is much easier to tell whether you convinced someone than it is to tell whether you beat them in an argument. Often, both parties think they won. In addition, sometimes it is rational not to update much in favor of a view if an advocate for that view beats you in an argument.

In support of this claim, consider what would happen if the world's smartest creationist debated some fairly ordinary evolution-believing high school student. The student would be destroyed in argument, but the student should not reject evolution, and I suspect he should hardly update at all. Why not? The student should know that there are people out there in the world who could destroy him on either side of this argument, and his personal ability to respond to arguments is not very relevant. What should be most relevant to this student is the distribution of opinion among people who are most trustworthy, not his personal response to small sample of the available evidence. Even if you genuinely are beating people in arguments, there is a risk that you will be like this creationist debater.

An additional consideration is that certain beliefs and practices may be reasonable and adopted for reasons that are not accessible to people who have adopted those beliefs and practices, as illustrated with the examples of the liver ritual and bedtime prayer. You might be able to "beat" some Christian in an argument about the merits of bedtime prayer, but praying may still be better than not praying. (I think it would be better still to introduce a different routine that serves similar functions—this is something I have done in my own life—but the Christian may be doing better than you on this issue if you don't have a replacement routine yourself.)

Under the elite common sense framework, the question is not "how reliable is elite common sense?" but "how reliable is elite common sense compared to me?" Suppose I learn that, actually, people are much worse at pricing derivatives than I previously believed. For the sake of argument suppose this was a lesson of the 2008 financial crisis (for the purposes of this argument, it doesn't matter whether this is actually a correct lesson of the crisis). This information does not favor relying more on my own judgment unless I have reason to think that the bias applies less to me than the rest of the derivatives market. By analogy, it is not acceptable to say, "People are really bad at thinking about philosophy. So I am going to give less weight to their judgments about philosophy (psst…and more weight to my personal hunches and the hunches of people I personally find impressive).” This is only OK if you have evidence that your personal hunches and the hunches of the people you personally find impressive are better than elite common sense, with respect to philosophy. In contrast, it might be acceptable to say, “People are very bad at thinking about the consequences of agricultural subsidies in comparison with economists, and most trustworthy people would agree with this if they had my evidence. And I have an unusual amount of information about what economists think. So my opinion gets more weight than elite common sense in this case.” Whether this ultimately is acceptable to say would depend on how good elites are at thinking about the consequences of agricultural subsidies—I suspect they are actually pretty good at it—but this is isn’t relevant to the general point that I’m making. The general point is that this is one potentially correct form of an argument that your opinion is better than the current stance of elite common sense.

This is partly a semantic issue, but I count the above example as a case where “you are more reliable than elite common sense,” even though, in some sense, you are relying on expert opinion rather than your own. But you have different beliefs about who is a relevant expert or what experts say than common sense does, and in this sense you are relying on your own opinion.

I favor giving more weight to common sense judgments in cases where people are trying to have accurate views. For example, I think people don’t try very hard to have correct political, religious, and philosophical views, but they do try to have correct views about how to do their job properly, how to keep their families happy, and how to impress their friends. In general, I expect people to try to have more accurate views in cases where it is in their present interests to have more accurate views. (A quick reference for this point is here.) This means that I expect them to strive more for accuracy in decision-relevant cases, cases where the cost of being wrong is high, and cases where striving for more accuracy can be expected to yield more accuracy, though not necessarily in cases where the risks and rewards are won’t come for a very long time. I suspect this is part of what explains why people can be skilled in tacit rationality but not explicit rationality.

As I said above, what’s critical is not how reliable elite common sense is but how reliable you are *in comparison with elite common sense*. So it only makes sense to give more weight to your views when learning that others aren’t trying to be correct if you have compelling evidence that you *are* trying to be correct. Ideally, this evidence would be compelling to a broad class of trustworthy people and not just compelling to you personally.

Some further reasons to think that the framework is likely to be helpful

In explaining the framework and outlining guidelines for applying it, I have given some reasons to expect this framework to be helpful. Here are some more weak arguments in favor of my view:

1. Some studies I haven’t personally reviewed closely claim that combinations of expert forecasts are hard to beat. For instance, a review by (Clemen 1989) found that: "Considerable literature has accumulated over the years regarding the combination of forecasts. The primary conclusion of this line of research is that

forecast accuracy can be substantially improved through the combination of multiple individual forecasts." (abstract) And a recent work by the Good Judgment Project found that taking an average individual forecasts and transforming it away from .5 credence gave the lowest errors of a variety of different methods of aggregating judgments of forecasters (p. 42).

2. There are plausible philosophical considerations suggesting that, absent special evidence, there is no compelling reason to favor your own epistemic standards over the epistemic standards that others use.
3. In practice, we are extremely reliant on conventional wisdom for almost everything we believe that isn't very closely related to our personal experience, and single individuals working in isolation have extremely limited ability to manipulate their environment in comparison with individuals who can build on the insights of others. To see this point, consider that a small group of very intelligent humans detached from all cultures wouldn't have much of an advantage at all over other animal species in competition for resources, but humans are increasingly dominating the biosphere. A great deal of this must be chalked up to cultural accumulation of highly adaptive concepts, ideas, and procedures that no individual could develop on their own. I see trying to rely on elite common sense as highly continuous with this successful endeavor.
4. Highly adaptive practices and assumptions are more likely to get copied and spread, and these practices and assumptions often work because they help you to be right. If you use elite common sense as a prior, you'll be more likely to be working with more adaptive practices and assumptions.
5. Some successful processes for finding valuable information, such as PageRank and Quora, seem analogous to the framework I have outlined. PageRank is one algorithm that Google uses to decide how high different pages should be in searches, which is implicitly a way of ranking high-quality information. I'm speaking about something I don't know very well, but my rough understanding is that PageRank gives pages more votes when more pages link to them, and votes from a page get more weight if that page itself has a lot of votes. This seems analogous to relying on elite common sense because information sources are favored when they are regarded as high quality by a broad coalition of other information sources. Quora seems analogous because it favors answers to questions that many people regard as good.
6. I'm going to go look at the first three questions I can find on Quora. I predict that I would prefer the answers that elite common sense would give to these questions to what ordinary common sense would say, and also that I would prefer elite common sense's answers to these questions to my own except in cases where I have strong inside information/analysis. Results: 1st question: weakly prefer elite common sense, don't have much special information. 2nd question: prefer elite common sense, don't have much special information. 3rd question: prefer elite common sense, don't have much special information. Note that I skipped a question because it was a matter of taste. This went essentially the way I predicted it to go.
7. The type of mathematical considerations underlying Condorcet's Jury Theorem give us some reason to think that combined opinions are often more reliable than individual opinions, even though the assumptions underlying this theorem are far from totally correct.
8. There's a general cluster of social science findings that goes under the heading "wisdom of crowds" and suggests that aggregating opinions across people outperforms individual opinions in many contexts.
9. Some rough "marketplace of ideas" arguments suggest that the best ideas will often become part of elite common sense. When claims are decision-relevant,

people pay if they have dumb beliefs and benefit if they have smart beliefs. When claims aren't decision-relevant, people sometimes pay a social cost for saying dumb things and get social benefits for saying things that are smarter, and the people with more information have more incentive to speak. For analogous reasons, when people use and promote epistemic standards that are dumb, they pay costs and when they use and promote epistemic standards that are smart. Obviously there are many other factors, including ones that point in different directions, but there is some kind of positive force here.

Cases where people often don't follow the framework but I think they should

I have seen a variety of cases where I believe people don't follow the principles I advocate. There are certain types of errors that I think many ordinary people make and others that are more common for sophisticated people to make. Most of these boil down to giving too much weight to personal judgments, giving too much weight to people who are impressive to you personally but not impressive by clear and uncontroversial standards, or not putting enough weight on what elite common sense has to say.

Giving too much weight to the opinions of people like you: People tend to hold religious views and political views that are similar to the views of their parents. Many of these people probably aren't trying to have accurate views. And the situation would be much better if people gave more weight to the aggregated opinion of a broader coalition of perspectives.

I think a different problem arises in the LessWrong and effective altruism communities. In this case, people are much more reflectively choosing which sets of people to get their beliefs from, and I believe they are getting beliefs from some pretty good people. However, taking an outside perspective, it seems overwhelmingly likely that these communities are subject to their own biases and blind spots, and the people who are most attracted to these communities are most likely to suffer from the same biases and blind spots. I suspect elite common sense would take these communities more seriously than it currently does if it had access to more information about the communities, but I don't think it would take us sufficiently seriously to justify having high confidence in many of our more unusual views.

Being overconfident on open questions where we don't have a lot of evidence to work with: In my experience, it is common to give little weight to common sense takes on questions about which there is no generally accepted answer, even when it is impossible to use commonsense reasoning to arrive at conclusions that get broad support. Some less sophisticated people seem to see this as a license to think whatever they want, as Paul Graham has commented in the case of politics and religion. I meet many more sophisticated people with unusual views about big picture philosophical, political, and economic questions in areas where they have very limited inside information and very limited information about the distribution of expert opinion. For example, I have now met a reasonably large number of non-experts who have very confident, detailed, unusual opinions about meta-ethics, libertarianism, and optimal methods of taxation. When I challenge people about this, I usually get some version of "people are not good at thinking about this question" but rarely a detailed explanation of why this person in particular is an exception to this generalization (more on this problem below).

There's an inverse version of this problem where people try to "suspend judgment" on questions where they don't have high-quality evidence, but actually end up taking very unusual stances without adequate justification. For example, I sometimes talk with people who say that improving the very long-term future would be overwhelmingly important if we could do it, but are skeptical about whether we can. In response, I sometimes run arguments of the form:

1. In expectation, it is possible to improve broad feature X of the world (education, governance quality, effectiveness of the scientific community, economic prosperity).
2. If we improve feature X, it will help future people deal with various big challenges and opportunities better in expectation.
3. If people deal with these challenges and opportunities better in expectation, the future will be better in expectation.
4. Therefore, it is possible to make the future better in expectation.

I've presented some preliminary thoughts on related issues here. Some people try to resist this argument on grounds of general skepticism about attempts at improving the world that haven't been documented with high-quality evidence. Peter Hurford's post on "speculative causes" is the closest example that I can point to online, though I'm not sure whether he still disagrees with me on this point. I believe that there can be some adjustment in the direction of skepticism in light of arguments that GiveWell has articulated here under "we are relatively skeptical," but I consider rejecting the second premise on these grounds a significant departure from elite common sense. I would have a similar view about anyone who rejected any of the other premises—at least if they rejected them for all values of X—for such reasons. It's not that I think the presumption in favor of elite common sense can't be overcome—I strongly favor thinking about such questions more carefully and am open to changing my mind—it's just that I don't think it can be overcome by these types of skeptical considerations. Why not? These types of considerations seem like they could make the probability distribution over impact on the very long-term narrower, but I don't see how they could put it tightly around zero. And in any case, GiveWell articulates other considerations in that post and other posts which point in favor of less skepticism about the second premise.

Part of the issue may be confusion about "rejecting" a premise and "suspending judgment." In my view, the question is "What are the expected long-term effects of improving factor X?" You can try not to think about this question or say "I don't know," but when you make decisions you are implicitly committed to certain ranges of expected values on these questions. To justifiably ignore very long-term considerations, I think you probably need your implicit range to be close to zero. I often see people who say they are "suspending judgment" about these issues or who say they "don't know" acting as if this ranger were very close to zero. I see this as a very strong, precise claim which is contrary to elite common sense, rather than an open-minded, "we'll wait until the evidence comes in" type of view to have. Another way to put it is that my claim that improving some broad factor X has good long-run consequences is much more of an anti-prediction than the claim that its expected effects are close to zero. (Independent point: I think that a more compelling argument than the argument that we can't affect the far future is the argument that that lots of ordinary actions have flow-through effects with astronomical expected impacts if anything does, so that people aiming explicitly at reducing astronomical waste are less privileged than one might think at first glance. I hope to write more about this issue in the future.)

Putting too much weight on your own opinions because you have better arguments on topics that interest you than other people, or the people you typically talk to: As mentioned above, I believe that some smart people, especially smart people who rely a lot on explicit reasoning, can become very good at developing strong arguments for their opinions without being very good at finding true beliefs. I think that in such instances, these people will generally not be very successful at getting a broad coalition of impressive people to accept their views (except perhaps by relying on non-rational methods of persuasion). Stress-testing your views by trying to actually convince others of your opinions, rather than just out-arguing them, can help you avoid this trap.

Putting too much weight on the opinions of single individuals who seem trustworthy to you personally but not to people in general, and have very unusual views: I have seen some people update significantly in favor of very unusual philosophical, scientific, and sociological claims when they encounter very intelligent advocates of these views. These people are often familiar with Aumann's agreement theorem and arguments for splitting the difference with epistemic peers, and they are rightly troubled by the fact that someone fairly similar to them disagrees with them on an issue, so they try to correct for their own potential failures of rationality by giving additional weight to the advocates of these very unusual views.

However, I believe that taking disagreement seriously favors giving these very unusual views less weight, not more. The problem partly arises because philosophical discussion of disagreement often focuses on the simple case of two people sharing their evidence and opinions with each other. But what's more relevant is the distribution of quality-weighted opinion around the world *in general*, not the distribution of quality-weighted opinion of the people that you have had discussions with, and not the distribution of quality-weighted opinion of the people that seem trustworthy to you personally. The epistemically modest move here is to try to stay closer to elite common sense, not to split the difference.

Objections to this approach

Objection: elite common sense is often wrong

One objection I often hear is that elite common sense is often wrong. I believe this is true, but not a problem for my framework. I make the comparative claim that elite common sense is more trustworthy than the idiosyncratic standards of the vast majority of individual people, not the claim that elite common sense is almost always right. A further consideration is that analogous objections to analogous views fail. For instance, "markets are often wrong in their valuation of assets" is not a good objection to the efficient markets hypothesis. As explained above, the argument that "markets are often wrong" needs to point to specific way in which one can do better than the market in order for it to make sense to place less weight on what the market says than on one's own judgments.

Objection: the best people are highly unconventional

Another objection I sometimes hear is that the most successful people often pay the least attention to conventional wisdom. I think this is true, but not a problem for my framework. One reason I believe this is that, according to my framework, when you go against elite common sense, what matters is whether elite common sense reasoning standards would justify your opinion if someone following those standards knew about your background, information, and analysis. Though I can't prove it, I suspect that the most successful people are often depart from elite common sense in ways that elite common sense would endorse if it had access to more information. I also believe that the most successful people tend to pay attention to elite common sense in many areas, and specifically bet against elite common sense in areas where they are most likely to be right.

A second consideration is that going against elite common sense may be a high-risk strategy, so that it is unsurprising if we see the most successful people pursuing it. People who give less weight to elite common sense are more likely to spend their time on pointless activities, join cults, and become crackpots, though they are also more likely to have revolutionary positive impacts. Consider an analogy: it may be that the gamblers who earned the most used the riskiest strategies, but this is not good evidence that you should use a risky strategy when gambling because the people who lost the most also played risky strategies.

A third consideration is that while it may be unreasonable to be too much of an independent thinker in a particular case, being an independent thinker helps you develop good epistemic habits. I think this point has a lot of merit, and could help explain why independent thinking is more common among the most successful people. This might seem like a good reason not to pay much attention to elite common sense. However, it seems to me that you can get the best of both worlds by being an independent thinker and keeping separate track of your own impressions and what elite common sense would make of your evidence. Where conflicts come up, you can try to use elite common sense to guide your decisions.

I feel my view is weakest in cases where there is a strong upside to disregarding elite common sense, there is little downside, and you'll find out whether your bet against conventional wisdom was right within a tolerable time limit. Perhaps many crazy-sounding entrepreneurial ideas and scientific hypotheses fit this description. I believe it may make sense to pick a relatively small number of these to bet on, even in cases where you can't convince elite common sense that you are on the right track. But I also believe that in cases where you really do have a great but unconventional idea, it will be possible to convince a reasonable chunk of elite common sense that your idea is worth trying out.

Objection: elite common sense is wrong about X, and can't be talked out of it, so your framework should be rejected in general

Another common objection takes the form: view X is true, but X is not a view which elite common sense would give much weight to. Eliezer makes a related argument here, though he is addressing a different kind of deference to common sense. He points to religious beliefs, beliefs about diet, and the rejection of cryonics as evidence that you shouldn't just follow what the majority believes. My position is closer to "follow the majority's epistemic standards" than "believe what the majority beliefs," and closer still to "follow the best people's epistemic standards without cherry picking "best" to suit your biases," but objections of this form could have some force against the framework I have defended.

A first response is that unless one thinks there are many values of X in different areas where my framework fails, providing a few counterexamples is not very strong evidence that the framework isn't helpful in many cases. This is a general issue in philosophy which I think is underappreciated, and I've made related arguments in chapter 2 of my dissertation. I think the most likely outcome of a careful version of this attack on my framework is that we identify some areas where the framework doesn't apply or has to be qualified.

But let's delve into the question about religion in greater detail. Yes, having some religious beliefs is generally more popular than being an atheist, and it would be hard to convince intelligent religious people to become atheists. However, my impression is that my framework does not recommend believing in God for the following reasons. Here are a number of weak arguments for this claim:

1. My impression is that the people who are most trustworthy by clear and generally accepted standards are significantly more likely to be atheists than the general population. One illustration of my perspective is that in a 1998 survey of the National Academy of Sciences, only 7% of respondents reported that they believed in God. However, there is a flame war and people have pushed many arguments on this issue, and scientists are probably unrepresentative of many trustworthy people in this respect.
2. While the world at large has broad agreement that some kind of higher power exists, there is very substantial disagreement about what this means, to the point where it isn't clear that these people are talking about the same thing.
3. In my experience, people generally do not try very hard to have accurate beliefs about religious questions and have little patience for people who want to carefully discuss arguments about religious questions at length. This makes it hard to stress-test one's views about religion by trying to get a broad coalition of impressive people to accept atheism, and makes it possible to give more weight to one's personal take if one has thought unusually carefully about religious questions.
4. People are generally raised in religious families, and there are substantial social incentives to remain religious. Social incentives for atheists to remain non-religious generally seem weaker, though they can also be substantial. For example, given my current social network, I believe I would pay a significant cost if I wanted to become religious.
5. Despite the above point, in my experience, it is much more common for religious people to become atheists than it is for atheists to become religious.
6. In my experience, among people who try very hard to have accurate beliefs about whether God exists, atheism is significantly more common than belief in God.
7. In my experience, the most impressive people who are religious tend not to behave much differently from atheists or have different takes on scientific questions/questions about the future.

These points rely a lot on my personal experience, could stand to be researched more carefully, and feel uncomfortably close to lousy contrarian excuses, but I think they are nevertheless suggestive. In light of these points, I think my framework recommends that the vast majority of people with religious beliefs should be substantially less confident in their views, recommends modesty for atheists who haven't tried very hard to be right, and I suspect it allows reasonably high confidence that God doesn't exist for people who have strong indicators that they have thought carefully about the issue. I think it would be better if I saw a clear and principled way for the framework to push more strongly in the direction of atheism, but the case has enough unusual features that I don't see this as a major argument against the general helpfulness of the framework.

As a more general point, the framework seems less helpful in the case of religion and politics because people are generally unwilling to carefully consider arguments with the goal of having accurate beliefs. By and large, when people are unwilling to carefully consider arguments with the goal of having accurate beliefs, this is evidence that it is not useful to try to think carefully about this area. This follows from the idea mentioned above that people tend to try to have accurate views when it is in their present interests to have accurate views. So if this is the main way the framework breaks down, then the framework is mostly breaking down in cases where good epistemology is relatively unimportant.

Conclusion

I've outlined a framework for taking account of the distribution of opinions and epistemic standards in the world and discussed some of its strengths and weaknesses. I think the largest strengths of the framework are that it can help you avoid falling prey to idiosyncratic personal biases, and that using it derives benefits from the "wisdom of crowds" effects. The framework is less helpful in:

1. cases where there is a large upside to disregarding elite common sense, there is little downside, and you'll find out whether your bet against conventional wisdom was right within a tolerable time limit, and
2. cases where people are unwilling to carefully consider arguments with the goal of having accurate beliefs.

Some questions for people who want to further develop the framework include:

1. How sensitive is the framework to other reasonable choices of standards for selecting trustworthy people? Are there more helpful standards to use?
2. How sensitive is the framework to reasonable choices of standards for aggregating opinions of trustworthy people?
3. What are the best ways of getting a better grip on elite common sense?
4. What other areas are there where the framework is particularly weak or particularly strong?
5. Can the framework be developed in ways that make it more helpful in cases where it is weakest?

Raising numerate children

How do you raise a child as a rationalist? I can't say that that was exactly what I had in mind but it seems to make for a fitting title here. A more precise title could have been: "How to deeply educate a child such that it fun and natural".

Today I'd like to tell you about the lullabies I sung and to what that led.

When my firstborn was very young I adapted a classic German lullaby "Schlaf kindlein schlaf" to numbers. It started with with only a few verses but grew over time (in part by the need to cover longer times until he slept).

I did sing it in German but I tried to translate it here to give you a better idea. It goes to the melody of "Schlaf Kindlein Schlaf which you may not know but can google easily (note that in German there are nicer rhymes for 100, million, googol):

> Sleep, baby, sleep!
>
> Thy father counts the sheep,
>
> One, two, three and four
>
> Little baby sleep some more
>
> Sleep, baby, sleep!

The refrain repeats and the verses are replaced as follows:

> Five, six, seven, eight - Tired at the dreamland gate.
>
> Nine, ten and eleven - Sleeping in the number heaven.
>
> Twelve, thirteen and fourteen - Sleeping babies have you seen.
>
> Fifteen up to twentyone - Dreaming baby sleep is done.
>
> Twentytwo to hundredtwo - Baby I will care for you.
>
> Hundredthree to thousendfive - Caring for you all your life.
>
> Thousandsix to millionthree - Of your dream you can break free.
>
> Millionfour to one googol - Dreaming of a giant ball.
>
> Googol-one to googolplex - Steaming in the dreamland tracks.
>
> Googolplex to infinity - I will always care for thee.

I get slower during the song and very slow with infinity - mostly they slept then.

I have dreams of this song where sheep accumulate to larger and larger blocks until the block number thats raising in blocks and everything ends in white noise.

I do no longer sing it to my older sons but they accompany me sometimes when singing it to my youngest (two years old). And they do know what googol means already.

I also have a bed ritual where I let them give the number of times I put the blanket on their face (they like it). When they give too large numbers I use blocks. These tended to get high too.

One time I asked for lower numbers (that was when my second oldest already knew halfs and quarters) which led to gaming for unusual fractions and ultimately to his insight that "There is no larger fraction than one half that can divide one" (by a seven year himself).

It seems to have put numbers so deeply in their mind and interest that my seven year old can do simple fractions, exponentials and roots in his head. I tried hard to avoid too much arithmetic before school lest they bore of math in school and that worked for his older brother (who nonetheless tops his class in math) but he just asks and asks and I have just given up and keep just answering his questions and posing comparable return questions at his Zone Of Proximal Development.

There are dialogs that run like this (contracted):

He: "In school we had to give tasks to get 50. I was allowed to give 5*10"

Me: "Can you give some other examples?"

He: "2*20+10" thinking a bit "20 time 2 and a half equals 50"

Me: "What about division?"

He: "100 divided by 2 obviously. Or 50/1."

Or:

Me: "How long is the side of a cube containing one litre?"

He: "10?" (omitting centimeters)

Me: "How do you know that?"

He: "You have told me." (*I* can't remember when; must be month's)

Me: "And how long is the side of a cube with 27 liters?"

He (dividing 27 then adding or something like that): "18,5?"

Me: "No. How did you get there?"

He: "There must be some number multiplied to get 27" (or something like that)

Me. "Yes, the side time the side times the side." (expecting him to try some numbers)

He: "What is the root of 27?" (he has picked up that root is the reverse of times the same number)

Me: "Good idea. Here whe have three times or a number to the 3rd power - so we need the 3rd root."

He: "And what is the third root or 27?"

Me: "Try it."

He: "2*2*2 equals 16 no 8" (he seems to remember a few powers of 2)

Me: "Yes. That is too small"

He: "5^3? 5*5*5?"

Me: "That is 125 - too large"

He: "3?"

Me: "Yes."

Or:

He: "What is 10*10*10*10*10?"

Me: "You mean 10 to the 5th power? That's hundred thousand"

He: "What 10*10*... [lots of 10s)?"

Me: "You mean 10 to the 30th power? Thats nonillion."

He: "What is 10 to the 100th power?"

Me: "That is called googol. A 1 with 100 zeros."

He: "What is 10^100^100^100?"

Me: "Do you mean 10^100 and that to the 100th power or 10 to the 100^100th power?"

He: Somehwat confused asks differnt questions, dialog levels off.

(note that in German "to the xth power" is simply "hoch" thus much easier to concatenate)

I have to say that I am quite proud of my children and wouldn't be surprised when you called me overly so. I have to add that we, my wife and I, invest significant time into our children, so just singing this song may not be enough. And it also may be that I was lucky that they are (partly) gifted with math (like me). But I have to emphasize that we did no rote memoization or repeated training whatsoever (and left that to school).

There are other things we do for 'rationalist training' which I will try to post some time soon.

Teaser:

- Bed time stories with complex patterns (endless stories, simply nested stories, parallel stories, forking stories).

- Everyday Experiments for young children.

Note 1: I place the lullaby under creative commons as checked below. You may adapt and I recommend finding better rhymes/meter.

Note 2: This is my first real post here. Please feel free to tell if you think it inappropriate, too long or too much showing my probable pride.

EDIT: fixed typos.

To what degree do you model people as agents?

The idea for this post came out of a conversation during one of the Less Wrong Ottawa events. A joke about being solipsist turned into a genuine question–if you wanted to assume that people were figments of your imagination, how much of a problem would this be? (Being told "you would be problematic if I were a solipsist" is a surprising compliment.)

You can rephrase the question as "do you model people as agents versus complex systems?" or "do you model people as PCs versus NPCs?" (To me these seem like a reframing of the same question, with a different connotation/focus; to other people they might seem like different questions entirely). Almost everyone at the table immediately recognized what we were talking about and agreed that modelling some people as agents and some people as complex systems was a thing they did. However, pretty much everything else varied–how much they modelled people as agents overall, how much it varied in between different people they knew, and how much this impacted the moral value that they assigned to other people. I suspect that another variable is "how much you model *yourself* as an agent"; this probably varies between people and impacts how they model others.

What does it mean to model someone as an agent?

The conversation didn't go here in huge amounts of detail, but I expect that due to typical mind fallacy, it's a fascinating discussion to have–that the distinctions that seem clear and self-evident to me probably aren't what other people use at all. I'll explain mine here.

1. *Reliability and responsibility*. Agenty people are people I feel I can rely on, who I trust to take heroic responsibility. If I have an unsolved problem and no idea what to do, I can go to them in tears and say "fix this please!" And they will do it. They'll pull out a solution that surprises me and that works. If the first solution doesn't work, they will keep trying.

In this sense, I model my parents strongly as agents–I have close to 100% confidence that they will do whatever it takes to solve a problem for me. There are other people who I trust to execute a pre-defined solution for me, once I've thought of it, like "could you do me a huge favour and drive me to the bike shop tomorrow at noon?" but whom I wouldn't go to with "AAAAH my bike is broken, help!" There are other people who I wouldn't ask for help, period. Some of them are people I get along with well and like a lot, but they aren't reliable, and they're further down the mental gradient towards NPC.

The end result of this is that I'm more likely to model people as agents if I know them well and have some kind of relationship where I would expect them to want to help me. Of course, this is incomplete, because there are brilliant, original people who I respect hugely, but who I don't know well, and I wouldn't ask or expect them to solve a problem in my day-to-day life. So this isn't the only factor.

2. *Intellectual formidability*. To what extent someone comes up with ideas that surprise me and seem like things I would never have thought of on my own. This also includes people who have accomplished things that I can't imagine myself succeeding

at, like startups. In this sense, there are a lot of bloggers, LW posters, and people on the CFAR mailing list who are major PCs in my mental classification system, but who I may not know personally at all.

3. *Conventional* "*agentiness*". The degree to which a person's behaviour can be described by "they wanted X, so they took action Y and got what they wanted", as opposed to "they did X kind of at random, and Y happened." When people seem highly agenty to me, I model their mental processes like this–my brother is one of them. I take the inside view, imagining that I wanted the thing they want and had their characteristics, i.e. relative intelligence, domain-specific expertise, social support, etc, and this gives better predictions than past behaviour. There are other people whose behaviour I predict based on how they've behaved in the past, using the outside view, while barely taking into account what they say they want in the future, and this is what gives useful predictions.

This category also includes the degree to which people have a growth mindset, which approximates how much they expect themselves to behave in an agenty way. My parents are a good example of people who are totally 100% reliable, but don't expect or want to change their attitudes or beliefs much in the next twenty years.

These three categories probably don't include all the subconscious criteria I use, but they're the main ones I can think of.

How does this affect relationships with people?

With people who I model as agents, I'm more likely to invoke phrases like "it was your fault that X happened" or "you said you would do Y, why didn't you?" The degree to which I feel blame or judgement towards people for not doing things they said they would do is almost directly proportional to how much I model them as agents. For people who I consider less agenty, whom I model more as complex systems, I'm more likely to skip the blaming step and jump right to "what are the things that made it hard for you to do Y? Can we fix them?"

On reflection, it seems like the latter is a healthier way to treat *myself*, and I know this (and consistently fail at doing this). However, I want to be treated like an agent by other people, not a complex system; I want people to give me the benefit of the doubt and assume that I know what I want and am capable of planning to get it. I'm not sure what this means for how I should treat other people.

How does this affect moral value judgements?

For me, not at all. My default, probably hammered in by years of nursing school, is to treat *every* human as worthy of dignity and respect. (On a gut level, it doesn't include animals, although it probably should. On an intellectual level, I don't think animals should be mistreated, but animal suffering doesn't upset me on the same visceral level that human suffering does. I think that on a gut level, my "circle of empathy" includes human dead bodies more than it includes animals).

One of my friends asked me recently if I got frustrated at work, taking care of people who had "brought their illness on themselves", i.e. by smoking, alcohol, drug use, eating junk food for 50 years, or whatever else people usually put in the category of "lifestyle choices." Honestly, I don't; it's not a distinction my brain makes. Some of my patients will recover, go home, and make heroic efforts to stay healthy; others won't, and will turn up back in the ICU at regular intervals. It doesn't affect how I feel about treating them; it feels meaningful either way. The one time I'm liable to get frustrated is when I have to spend hours of hard work on patients who are severely neurologically damaged and are, in a sense, dead already, or at least not people anymore. I hate this. But my default is still to talk to them, keep them looking tidy and comfortable, et cetera...

In that sense, I don't know if modelling different people differently is, for me, a morally a right or a wrong thing to do. However, I spoke to someone whose default is *not* to assign people moral value, unless he models them as agents. I can see this being problematic, since it's a high standard.

Conclusion

As usual for when I notice something new about my thinking, I expect to pay a *lot* of attention to this over the next few weeks, and probably notice some interesting things, and quite possibly change the way I think and behave. I think I've already succeeded in finding the source of some mysterious frustration with my roommate; I want to model her as an agent because of #1–she's my best friend and we've been through a lot together–but in the sense of #3, she's one of the least agenty people I know. So I consistently, *predictably* get mad at her for things like saying she'll do the dishes and then not doing them, and getting mad doesn't help either of us at all.

I'm curious to hear what other people think of this idea.

New Monthly Thread: Bragging

In an attempt to encourage more people to *actually do awesome things* (a la instrumental rationality), I am proposing a new monthly thread (can be changed to bi-weekly, should that be demanded). Your job, should you choose to accept it, is to comment on this thread explaining **the most awesome thing you've done this month**. You may be as blatantly proud of you self as you feel. You may unabashedly consider yourself *the coolest freaking person ever* because of that awesome thing you're dying to tell everyone about. This is the place to do just that.

Remember, however, that this **isn't** any kind of progress thread. Nor is it any kind of proposal thread.*This thread is solely for people to talk about the awesomest thing they've done all month. not will do. not are working on*. **have already done.** This is to cultivate an environment of object level productivity rather than meta-productivity methods.

So, what's the coolest thing you've done this month?