Best of LessWrong: March 2012

Best of LessWrong: March 2012

Schelling fences on slippery slopes

Slippery slopes are themselves a slippery concept. Imagine trying to explain them to an alien:

"Well, we right-thinking people are quite sure that the Holocaust happened, so banning Holocaust denial would shut up some crackpots and improve the discourse. But it's one step on the road to things like banning unpopular political positions or religions, and we right-thinking people oppose that, so we won't ban Holocaust denial."

And the alien might well respond: "But you could just ban Holocaust denial, but not ban unpopular political positions or religions. Then you right-thinking people get the thing you want, but not the thing you don't want."

This post is about some of the replies you might give the alien.

Abandoning the Power of Choice

This is the boring one without any philosophical insight that gets mentioned only for completeness' sake. In this reply, giving up a certain point risks losing the ability to decide whether or not to give up other points.

For example, if people gave up the right to privacy and allowed the government to monitor all phone calls, online communications, and public places, then if someone launched a military coup, it would be very difficult to resist them because there would be no way to secretly organize a rebellion. This is also brought up in arguments about gun control a lot.

I'm not sure this is properly thought of as a slippery slope argument at all. It seems to be a more straightforward "Don't give up useful tools for fighting tyranny" argument.

The Legend of Murder-Gandhi

Previously on Less Wrong's *The Adventures of Murder-Gandhi*: Gandhi is offered a pill that will turn him into an unstoppable murderer. He refuses to take it, because in his current incarnation as a pacifist, he doesn't want others to die, and he knows that would be a consequence of taking the pill. Even if we offered him $1 million to take the pill, his abhorrence of violence would lead him to refuse.

But suppose we offered Gandhi $1 million to take a different pill: one which would decrease his reluctance to murder by 1%. This sounds like a pretty good deal. Even a person with 1% less reluctance to murder than Gandhi is still pretty pacifist and not likely to go killing anybody. And he could donate the money to his favorite charity and perhaps save some lives. Gandhi accepts the offer.

Now we iterate the process: every time Gandhi takes the 1%-more-likely-to-murder-pill, we offer him another $1 million to take the same pill again.

Maybe original Gandhi, upon sober contemplation, would decide to accept $5 million to become 5% less reluctant to murder. Maybe 95% of his original pacifism is the only level at which he can be *absolutely sure* that he will still pursue his pacifist ideals.

Unfortunately, original Gandhi isn't the one making the choice of whether or not to take the 6th pill. 95%-Gandhi is. And 95% Gandhi doesn't care *quite* as much about pacifism as original Gandhi did. He still doesn't want to become a murderer, but it wouldn't be a disaster if he were just 90% as reluctant as original Gandhi, that stuck-up goody-goody.

What if there were a general principle that each Gandhi was comfortable with Gandhis 5% more murderous than himself, but no more? Original Gandhi would start taking the pills, hoping to get down to 95%, but 95%-Gandhi would start taking five more, hoping to get down to 90%, and so on until he's rampaging through the streets of Delhi, killing everything in sight.

Now we're tempted to say Gandhi shouldn't even take the first pill. But this also seems odd. Are we really saying Gandhi shouldn't take what's basically a free million dollars to turn himself into 99%-Gandhi, who might well be nearly indistinguishable in his actions from the original?

Maybe Gandhi's best option is to "fence off" an area of the slippery slope by establishing a Schelling point - an arbitrary point that takes on special value as a dividing line. If he can hold himself to the precommitment, he can maximize his winnings. For example, original Gandhi could swear a mighty oath to take only five pills - or if he didn't trust even his own legendary virtue, he could give all his most valuable possessions to a friend and tell the friend to destroy them if he took more than five pills. This would commit his future self to stick to the 95% boundary (even though that future self is itching to try to the same precommitment strategy to stick to its own 90% boundary).

Real slippery slopes will resemble this example if, each time we change the rules, we also end up changing our opinion about how the rules should be changed. For example, I think the Catholic Church may be working off a theory of "If we give up this traditional practice, people will lose respect for tradition and want to give up even more traditional practices, and so on."

Slippery Hyperbolic Discounting

One evening, I start playing *Sid Meier's Civilization* (IV, if you're wondering - V is terrible). I have work tomorrow, so I want to stop and go to sleep by midnight.

At midnight, I consider my alternatives. For the moment, I feel an urge to keep playing Civilization. But I know I'll be miserable tomorrow if I haven't gotten enough sleep. Being a hyperbolic discounter, I value the next ten minutes a lot, but after that the curve becomes pretty flat and maybe I don't value 12:20 much more than I value the next morning at work. Ten minutes' sleep here or there doesn't make any difference. So I say: "I will play Civilization for ten minutes - 'just one more turn' - and then I will go to bed."

Time passes. It is now 12:10. Still being a hyperbolic discounter, I value the next ten minutes a lot, and subsequent times much less. And so I say: I will play until 12:20, ten minutes sleep here or there not making much difference, and then sleep.

And so on until my empire bestrides the globe and the rising sun peeps through my windows.

This is pretty much the same process described above with Murder-Gandhi except that here the role of the value-changing pill is played by time and my own tendency to discount hyperbolically.

The solution is the same. If I consider the problem early in the evening, I can precommit to midnight as a nice round number that makes a good Schelling point. Then, when deciding whether or not to play after midnight, I can treat my decision not as "Midnight or 12:10" - because 12:10 will always win *that* particular race - but as "Midnight or abandoning the only credible Schelling point and probably playing all night", which will be sufficient to scare me into turning off the computer.

(if I consider the problem at 12:01, I may be able to precommit to 12:10 if I am especially good at precommitments, but it's not a very natural Schelling point and it might be easier to say something like "as soon as I finish this turn" or "as soon as I discover this technology").

Coalitions of Resistance

Suppose you are a Zoroastrian, along with 1% of the population. In fact, along with Zoroastrianism your country has fifty other small religions, each with 1% of the population. 49% of your countrymen are atheist, and hate religion with a passion.

You hear that the government is considering banning the Taoists, who comprise 1% of the population. You've never liked the Taoists, vile doubters of the light of Ahura Mazda that they are, so you go along with this. When you hear the government wants to ban the Sikhs and Jains, you take the same tack.

But now you are in the unfortunate situation described by Martin Niemoller:

> *First they came for the socialists, and I did not speak out, because I was not a socialist.*
> *Then they came for the trade unionists, and I did not speak out, because I was not a trade unionist.*
> *Then they came for the Jews, and I did not speak out, because I was not a Jew.*
> *Then they came for me, but we had already abandoned the only defensible Schelling point*

With the banned Taoists, Sikhs, and Jains no longer invested in the outcome, the 49% atheist population has enough clout to ban Zoroastrianism and anyone else they want to ban. The better strategy would have been to have all fifty-one small religions form a coalition to defend one another's right to exist. In this toy model, they could have done so in an ecumenial congress, or some other literal strategy meeting.

But in the real world, there aren't fifty-one well-delineated religions. There are billions of people, each with their own set of opinions to defend. It would be impractical for everyone to physically coordinate, so they have to rely on Schelling points.

In the original example with the alien, I cheated by using the phrase "right-thinking people". In reality, figuring out who qualifies to join the Right-Thinking People Club is half the battle, and everyone's likely to have a different opinion on it. So far, the practical solution to the coordination problem, the "only defensible Schelling point", has been to just have everyone agree to defend everyone else without worrying whether they're right-thinking or not, and this is easier than trying to coordinate room for exceptions like Holocaust deniers. Give up on the Holocaust deniers, and no one else can be sure what other Schelling point you've committed to, if any...

...unless they can. In parts of Europe, they've banned Holocaust denial for years and everyone's been totally okay with it. There are also a host of other well-respected exceptions to free speech, like shouting "fire" in a crowded theater. Presumably, these exemptions are protected by tradition, so that they have become new Schelling points there, or are else so obvious that everyone except Holocaust deniers is willing to allow a special Holocaust denial exception without worrying it will impact their own case.

Summary

Slippery slopes legitimately exist wherever a policy not only affects the world directly, but affects people's willingness or ability to oppose future policies. Slippery slopes can sometimes be avoided by establishing a "Schelling fence" - a Schelling point that the various interest groups involved - or yourself across different values and times - make a credible precommitment to defend.

Using degrees of freedom to change the past for fun and profit

Follow-up to: Follow-up on ESP study: "We don't publish replications", Feed the Spinoff Heuristic!

Related to: Parapsychology: the control group for science, Dealing with the high quantity of scientific error in medicine

> Using the same method as in Study 1, we asked 20 University of Pennsylvania undergraduates to listen to either “When I’m Sixty-Four” by The Beatles or “Kalimba.” Then, in an ostensibly unrelated task, they indicated their birth date (mm/dd/yyyy) and their father’s age. We used father’s age to control for variation in baseline age across participants. An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to “When I’m Sixty-Four” (adjusted M = 20.1 years) rather than to “Kalimba” (adjusted M = 21.5 years), F(1, 17) = 4.92, p = .040

That's from "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," which runs simulations of a version of Shalizi's "neutral model of inquiry," with random (null) experimental results, augmented with a handful of choices in the setup and analysis of an experiment. Even before accounting for publication bias, these few choices produced a desired result "significant at the 5% level" 60.7% of the time, and at the 1% level 21.5% at the time.

I found it because of another paper claiming time-defying effects, during a search through all of the papers on Google Scholar citing Daryl Bem's precognition paper, which I discussed in a past post about the problems of publication bias and selection over the course of a study. For Bem, Richard Wiseman established a registry for the methods, and tests of the registered studies could be set prior to seeing the data (in addition to avoiding the file drawer).

Now a number of purported replications have been completed, with several available as preprints online, including a large "straight replication" carefully following the methods in Bem's paper, with some interesting findings discussed below. The picture does not look good for psi, and is a good reminder of the sheer cumulative power of applying a biased filter to many small choices.

Background

When Bem's article was published the skeptic David Alcock argued that Bem's experiments involved midstream changes of methods, choices in the transformation of data (raw data was not available), and other signs of modifying the experiment and analysis in response to the data. Wagenmakers et al drew attention to writing by Bem advising young psychologists to take experiments that failed to show predicted effects and relentlessly explore the data in hopes of generating an attractive and significant effect. In my post, I emphasized the importance of "straight replications," with methodology, analytical tests, and intent to publish established in advance, as in Richard Wiseman's registry of studies.

An article by Gregory Francis uses a standard test for publication bias on Bem's article: comparing the number of findings reaching significance to the number predicted by the power of the study to detect the claimed effect. 9 of the 10 experiments mentioned described in Bem's article[1] find positive effects using Bem's measures and tests, but those 9 were all statistically significant despite the small size of the effects. Francis calculates a 5.8% probability of so many reaching significance by chance (given the estimated effect power and effect size).

Other complaints included declining effect size with sample size (driven mostly by one larger experiment), the use of one-tailed tests (Bem justified this as following an early hypothesis, but claims of "psi-missing" due to boredom or repelling stimuli are found in the literature and could have been mustered), and the failure to directly replicate a single experiment or concentrate subjects.

Subsequent replications

At the time of my first post, I was able to find several replication attempts already online. Richard Wiseman and his coauthors had not found psi, and were refused consideration for publication at the journal which had hosted the original article. Galak and Nelson had tried and failed to replicate experiment 8. A a pro-psi researcher had pulled a different 2006 experiment from the file drawer and retitled as a purported "replication" of the 2011 paper. Samuel Moulton, who previously worked with Bem, writes that he tried to replicate Bem with 200 subjects and found no effect (not just not a significant effect, but a significantly lower effect), but that Bem would not mention this in the 2011 publication. Bem confirms this in a video of a Harvard debate.

Since then, there have been more replications. This New Scientist article claims to have found 7 replications of Bem, with six failures and one success. The success is said to be by a researcher who has previously studied the effect of "geomagnetic pulsations" on ESP, but I could not locate it online.

Snodgrass (2011) failed to replicate Bem using a version of the Galak and Nelson experiment. Wagenmaker et al posted their methods in advance, but have not yet posted their results, although news media have reported that they also got a negative result Bem. Wiseman and his coauthors posted their abstract online, and claim to have performed a close replication of one of Bem's experiments with three times the subjects, finding no effect (despite 99%+ power to detect Bem's claimed effect). Another paper, "Correcting the Past: Failures to Replicate Psi," by Galak, LeBoeuf, Nelson, and Simmons, combines 6 experiments by the researchers (who are at four separate universities) with 820 subjects and finds no effect in a very straight replication. More on it in a moment.

I also found the abstracts of the 2011 Towards a Science of Consciousness conference. On page 166 Whitmarsh and Bierman claim to have conducted a replication of a Bem experiment involving meditators, but do not give their results, although it appears they may have looked for effects of meditation on the results. On page 176, there is an abstract from Franklin and Schooler, claiming success in a new and different precognition experiment, as well as predicting the outcome of a roulette wheel (n=204, hit rate 57%, p<.05). In the New Scientist article they claim to have replicated their experiment (with much reduced effect size and just barely above the 0.05 significance level), although past efforts to use psi in casino games have not been repeatable (nor have the experimenters become mysteriously wealthy, or easily able to fund their research, apparently). The move to a new and ill-described format prevents it from being used as a straight replication (in Shalizi's neutral model of inquiry using only publication bias, it is the move to new effects lets a field sustain itself in the absence of a subject matter), it was not registered, and the actual study is not available, so I will leave it be until publication.

Correcting the Past: Failures to Replicate Psi

Throughout this paper the researchers try to specify their procedures unambigously and as closely aligned with Bem as they can, for instance in transforming the data[2] so as to avoid cherry-picking in the fashion they argue:

> Results
>
> To test for the presence of precognition, Bem (2011) computed a weighted differential recall score (DR) for each participant using the formula:
>
> DR = (Recalled Practiced Words - Recalled Control Words) ×
>
> (Recalled Practice Words + Recalled Control Words)
>
> In the paper, for descriptive purposes, Bem frequently reports this number as DR%, which is the percentage that a participant's score deviated from random chance towards the highest or lowest scores possible (-576 to 576). We conducted the identical analysis on our data and also report DR% (see Table 1). In addition to using the weighted differential recall score, we also report the results using a simple unweighted recall score, which is the difference between recalled practice words and recalled control words (see Appendix B). For both of these measures, random chance would lead to a score of 0, and analysis was conducted using a one-sample t-test.

This prevents them from choosing the more favorable (or less favorable) of several transformations, as they seem to suggest Bem did in the next quote, bumping a result to significance in the original paper. This is a recurrent problem across many fields, and a reason to seek out raw data whenever possible, or datasets collected by neutral parties (on your question of interest):

> Still, even in Experiments 8 and 9, it is unclear how Bem could find significant support for a hypothesis that appears to be untrue. Elsewhere, critics of Bem have implicated his use of a one-tailed statistical test (Wagenmakers et al. 2011), testing multiple comparisons without correction (Wagenmakers et al. 2011), or perhaps simply a lurking file drawer with some less successful pilot experiments. All of these concerns fall under a larger category of researcher degrees of freedom, which raise the likelihood of falsely rejecting the null hypothesis (Simmons et al., 2011). Some of these can be easily justifiable and have small and seemingly inconsequential effects. For example, Bem analyzes participant recall using an algorithm which weights the total number of correctly recalled words (i.e., DR%). He could, presumably, have just as easily analyzed simple difference scores and found a similar, but not quite identical, result (indeed, re-analyzing the data from Bem (2011) Experiment 8 with a simple difference score yields no Psi effects (M = .49, t(99) = 1.48, p = .14), though it does for Experiment 9 (M =.96; t(49) = 2.46, p = .02)).

They mention others which they did not have data to test:

> The scoring distinction is just a single example, but even for Bem's simple procedure there are many others. For example, Bem's words are evenly split between common and uncommon words, a difference that was not analyzed (or

> reported) in the original paper, but may reflect an alternative way to consider the data (perhaps psi only persists for uncommon words? Perhaps only for common words?). He reports the results of his two-item sensation seeking measure, but he does not analyze (or report collecting) additional measures of participant anxiety or experimenter-judged participant enthusiasm. Presumably these were collected because there was a possibility that they may be influential as well, but when analysis revealed that they were not, that analysis was dropped from the paper.

Other elements providing degrees of freedom were left out of the Bem paper. A published paper can only provide so much confidence that it actually describes the experiment as it happened (or didn't!):

> Despite our best efforts to conduct identical replications of Bem’s Experiments 8 and 9, it is possible that the detection of psi requires certain methodological idiosyncrasies that we failed to incorporate into our experiments. For instance, after reading the replication packet (personal communication with Bem, November, 1 2010) provided by Bem, we noticed that there were at least three differences between our experiments (which mirrored the exact procedure described in Bem’s published paper) and the procedure actually employed by Bem...the experimenter was required to have a conversation with each participant in order to relax him or her...participants were asked two questions in addition to the sensation seeking scale...the set of words used by Bem were divided into common and uncommon words, something that we did not do in our Experiments 1 and 2.

The experiments, with several times the collective sample size of the Bem experiments (8 and 9) they replicate, look like chance:

> Main Results
>
> Table 1 presents the results of our six experiments as well as the results from Bem’s (2011) Experiments 8 and 9, for comparison. Bem found DR% = 2.27% in Experiment 8 and 4.21% in Experiment 9, effects that were significant at p = .03 and p = .002, one-tailed.
>
> In contrast, none of our six experiments showed a significant effect suggesting precognition.
>
> In Experiment 1, DR% = -1.21%, t(111) = -1.201, p = .23 (all p-values in this paper are two-tailed). Bayesian t-tests (advocated by Wagenmakers et al., 2011) suggest that this is “substantial” support for the null hypothesis of no precognition.
>
> In Experiment 2, DR% = 0.00%, t(157) = .00, p = .99. Bayesian t-tests suggest that this is “strong” support for the null hypothesis.
>
> In Experiment 3, DR% = 1.17%, t(123) = 1.28, p = .20. Although DR% was indeed above zero, in the direction predicted by the ESP hypothesis, the test statistic did not reach conventional levels of significance, and Bayesian t-tests suggest that this is nevertheless “substantial” support for the null hypothesis.
>
> In Experiment 4, DR% = 1.59%, t(108) = 1.77, p = .08. Again, although DR% was above zero, the test statistic did not reach conventional levels of significance, and Bayesian t-tests still suggest that this is “substantial” support for the null hypothesis.

> In Experiment 5, which contained our largest sample of participants, DR% = -.49%, t(210) = -.71, p = .48. Bayesian t-tests suggest that this is “strong” support for the null hypothesis.
>
> Finally, in Experiment 6’s Test-Before-Practice condition, DR% = -.29%, t(105) = -.33, p = .74. Bayesian t-tests suggest that this is “strong” support for the null hypothesis.
>
> In sum, in four of our experiments, participants recalled more control words than practice words (Experiments 1, 2, 5, and 6) and in two of our experiments, participants recalled more practice words than control words (Experiments 3 and 4). None of these effects were statistically reliable using conventional t-tests (see Table 1). As noted, Bayesian t-tests suggest that even the two findings that were directionally consistent with precognition show substantial support for the null hypothesis of no precognition.

Perhaps the reported positive replication will hold up to scrutiny (with respect to sample size, power, closeness of replication, data mining, etc), or some other straight replication will come out convincingly positive (in light of the aggregate evidence). I doubt it.

Psi and science

Beating up on parapsychology may be cheap and easy in the scientific, skeptical, and Less Wrong communities, a low-status outgroup belief. But the abuse of many degrees of freedom, and shortage of close replication, is widespread in science and particularly in psychology. The heuristics and biases literature, studies of cognitive enhancement, social psychology and other areas often used in Less Wrong are not so different. This suggests a candidate hack to fight confirmation bias in assessing the evidentiary value of experiments that confirm one's views: ask yourself how much evidentiary weight (in log odds) you would place on the same methods and results showing a novel psi effect?[3]

Notes

[1] In addition to the nine numbered experiments, there is a footnote referring to a small early tenth study which did not find an effect in Bem 2011.

[2] One of the bigger differences is that some of the experiments were online rather than in the lab, but this didn't seem to matter much. They also switched from blind human coding of misspelled words to computerized coding.

[3] This heuristic has not been tested, beyond the general (psychology!) results suggesting that arguing for a position opposite your own can help to see otherwise selectively missed considerations.

ETA: This blog post also discusses the signs of optional stopping, multiple hypothesis testing, use of one-tailed tests where a negative result could also have been reported as due to psi, etc.

ETA2: A post at the Bare Normality blog tracks down earlier presentation of some of the experiments going into Bem (2011), back in 2003, and notes that the data seem to bee selectively ported to the 2011 paper, described quite differently, and discusses other signs of unreported experiments. The post also expresses concern about reconciling these data with Bem's explicit denial of optional stopping, selective reporting, and similar.

ETA3: Bem's paper cites an experiment by Savva as evidence for precognition (by arachnophobes), but leaves out the fact that Savva's follow-up experiments failed to replicate the effect. Links and references are provided in a post at the James Randi forums. Savva also says that Bem had "extracted" several supposedly significant precognition correlations from Savva's data, and upon checking Savva found they were generated by calculation errors. Bem also is said to have claimed Savva's first result had passed the 0.05 significance test, when it was actually just short of doing so (0.051, not a substantial difference, and perhaps defensible, but another sign of bias).

Is community-collaborative article production possible?

When I showed up at the Singularity Institute, I was surprised to find that 30-60 papers' worth of material was lying around in blog posts, mailing list discussions, and people's heads — but it had never been written up in clear, well-referenced academic articles.

Why is this so? Writing such articles has many clear benefits:

- Clearly stated and well-defended arguments can persuade smart people to take AI risk seriously, creating additional supporters and collaborators for the Singularity Institute.
- Such articles can also improve the credibility of the organization as a whole, which is especially important for attracting funds from top-level social entrepreneurs and institutions like the Gates Foundation and Givewell.
- Laying out the arguments clearly and analyzing each premise can lead to new strategic insights that will help us understand how to purchase x-risk reduction most efficiently.
- Clear explanations can provide a platform on which researchers can build to produce new strategic and technical research results.
- Communicating clearly is what lets other people find errors in your reasoning.
- Communities can use articles to cut down on communication costs. When something is written up clearly, 1000 people can read a single article instead of needing to transmit the information by having several hundred personal conversations between 2-5 people.

Of course, there are costs to writing articles, too. The single biggest cost is *staff time / opportunity cost*. An article like "Intelligence Explosion: Evidence and Import" can require anywhere from 150-800 person-hours. That is 150-800 paid hours during which our staff is *not* doing other critically important things that collectively have a bigger positive impact than a single academic article is likely to have.

So Louie Helm and Nick Beckstead and I sat down and asked, "Is there a way we can buy these articles without such an egregious cost?"

We think there might be. Basically, we suspect that most of the work involved in writing these articles can be outsourced. Here's the process we have in mind:

1. An SI staff member chooses a paper idea we need written up, then writes an abstract and some notes on the desired final content.
2. SI pays Gwern or another remote researcher to do a literature search-and-summary of relevant material, with pointers to other resources.
3. SI posts a **contest** to LessWrong, inviting submissions of near-conference-level-quality articles that follow the provided abstract and notes on desired final content. Contestants benefit by starting with the results of Gwern's literature summary, and by knowing that they don't need to produce something as good as "Intelligence Explosion: Evidence and Import" to win the prize. First place wins $1200, 2nd place wins $500, and 3rd place wins $200.
4. Submissions are due 1 month later. Submission are reviewed, and the authors of the best submissions are sent comments on what could be improved to

maximize the chances of coming in first place.

5. Revised articles are due 3 weeks after comments are received. Prizes are awarded.
6. SI pays an experienced writer like Yvain or Kaj_Sotala or someone similar to build up and improve the 1st place submission, borrowing the best parts from the other submissions, too.
7. An SI staff member does a final pass, adding some content, making it more clearly organized and polished, etc. One of SI's remote editors does another pass to make the sentences more perfect.
8. The paper is submitted to a journal or an edited volume, and is marked as being co-authored by (1) the key SI staff member who provided the seed ideas and guided each stage of the revisions and polishing, (2) the author of the winning submission, and (3) Gwern. (With thanks to contributions from the other contest participants whose submissions were borrowed from — unless huge pieces were borrowed, in which case they may be counted as an additional co-author.)

If this method works, each paper may require only 50-150 hours of SI staff time per paper — a dramatic improvement! But this method has additional benefits:

- Members of the community who are capable of doing one piece of the process but not the other pieces get to contribute where they shine. (Many people can write okay-level articles but can't do efficient literature searches or produce polished prose, etc.)
- SI gets to learn more about the talent that exists in its community which hadn't yet been given the opportunity to flower. (We might be able to directly outsource future work to contest participants, and if one person wins three such contests, that's an indicator that we should consider hiring them.)
- Additional paid "jobs" (by way of contest money) are created for LW rationalists who have some domain expertise in singularity-related subjects.
- Many Less Wrongers are students in fields relevant to the subject matter of the papers that will be produced by this process, and this will give them an opportunity to co-author papers that can go on their CV.
- The community in general gets better at collaborating.

This is, after all, more similar to how many papers would be produced by university departments, in which a senior researcher works with a team of students to produce papers.

Feedback? Interest?

(Not exactly the same, but see also the Polymath Project.)

Main section vs. discussion section

(The following may only apply to me. I mention it to see if anyone else has had the same issue).

For a long time I have been only looking at the Discussion section and promoted main page articles. Just now on a whim I checked the non-promoted main page articles and found there were a whole bunch of them, some potentially quite interesting, that I had missed. My expectation based on past experience was that all reasonably good articles from main would be "promoted", but perhaps this has changed. If this has been going on for a while I've presumably missed quite a bit of content. Perhaps it should be made easier to find/notice these? It's a bit weird and awkward that there are 3 different non-uniform ways of finding posts.

New cognitive bias articles on wikipedia (update)

- Conservatism
- Curse of knowledge
- Duration neglect
- Extension neglect
- Extrinsic incentives bias
- Illusion of external agency
- Illusion of validity
- Insensitivity to sample size
- Lady Macbeth effect
- Less-is-better effect
- Naïve cynicism
- Naïve realism
- Reactive devaluation
- Rhyme-as-reason effect
- Scope neglect

Also conjunction fallacy has been expanded.

(update) background
I started dozens of the cognitive bias articles that are on wikipedia. That was a long time ago. It seems people like these things, so I started adding them again.
I wanted to write a compendium of biases in book form. I didn't know how to get a book published, though.
Anyway, enjoy.

How to avoid dying in a car crash

Aside from cryonics and eating better, what else can we do to live long lives?

Using this tool, I looked up the risks of death for my demographic group. As a 15-24 year old male in the United States, the most likely cause of my death is a traffic accident; and so I'm taking steps to avoid that. Below I have included the results of my research as well as the actions I will take to implement my findings. Perhaps my research can help you as well.[1]

Before diving into the results, I will note that this data took me *one hour* to collect. It's definitely not comprehensive, and I know that working together, we can do much better. So if you have other resources or data-backed recommendations on how to avoid dying in a traffic accident**, leave a comment below and I'll update this post.**

General points

Changing your behavior *can* reduce your risk of death in a car crash. A 1985 report on British and American crash data discovered that "driver error, intoxication and other human factors contribute wholly or partly to about 93% of crashes." Other drivers' behavior matters too, of course, but you might as well optimize your own.[2]

Secondly, **overconfidence appears to be a large factor in peoples' thinking about traffic safety.** A speaker for the National Highway Traffic Safety Association (NHTSA) stated that "Ninety-five percent of crashes are caused by human error… but 75% of drivers say they're more careful than most other drivers. Less extreme evidence for overconfidence about driving is presented here.

One possible cause for this was suggested by the Transport Research Laboratory, which explains that "...the feeling of being confident in more and more challenging situations is experienced as evidence of driving ability, and that 'proven' ability reinforces the feelings of confidence. Confidence feeds itself and grows unchecked until something happens – a near-miss or an accident."

So if you're tempted to use this post as an opportunity to feel superior to *other* drivers, remember: you're probably overconfident too! Don't just humbly confess your imperfections – change your behavior.

Top causes of accidents

Distraction

Driver distraction is one of the largest causes of traffic accident deaths. The Director of Traffic Safety at the American Automobile Association stated that "The research tells us that somewhere between 25-50 percent of all motor vehicle crashes in this country really have driver distraction as their root cause." The NHTSA reports the number as 16%.

If we are to reduce distractions while driving, we ought to identify which distractors are the worst. One is cell phone use. My solution: Don't make calls in the car, and turn off your phone's sound so that you aren’t tempted.

I brainstormed other major distractors and thought of ways to reduce their distracting effects.

Distractor: Looking at directions on my phone as I drive

- Solution: Download a great turn-by-turn navigation app (recommendations are welcome).
- Solution: Buy a GPS.

Distractor: Texting, Facebook, slowing down to gawk at an accident, looking at scenery

- Solution [For System 2]: Consciously accept that texting (Facebook, gawking, scenery) causes accidents.
- Solution [For System 1]: Once a week, vividly and emotionally imagine texting (using Facebook, gawking at an accident) and then crashing & dying.
- Solution: Turn off your phone’s sound while driving, so you won’t answer texts.

Distractor: Fatigue

- Solution [For System 2]: Ask yourself if you’re tired before you plan to get in the car. Use Anki or a weekly review list to remember the association.
- Solution [For System 1]: Once a week, vividly and emotionally imagine dozing off while driving and then dying.

Distractor: Other passengers

- Solution: Develop an identity as someone who drives safely and thinks it’s low status to be distracting in the car. Achieve this by meditating on the commitment, writing a journal entry about it, using Anki, or saying it every day when you wake up in the morning.
- Solution [In the moment]: Tell people to chill out while you’re driving. Mentally simulate doing this ahead of time, so you don’t hesitate to do it when it matters.

Distractor: Adjusting the radio

- Solution: If avoiding using the car radio is unrealistic, minimize your interaction with it by only using the hotkey buttons rather than manually searching through channels.
- Solution: If you’re constantly tempted to change the channel (like I am), buy an iPod cable so you can listen to your own music and set playlists that you like, so you won't constantly want to change the song.

A last interesting fact about distraction, from Wikipedia:

> Recent research conducted by British scientists suggests that music can also have an effect [on driving]; classical music is considered to be calming, yet too much could relax the driver to a condition of distraction. On the other hand, hard rock may encourage the driver to step on the acceleration pedal, thus creating a potentially dangerous situation on the road.

Speeding

The Road and Traffic Authority of New South Wales claims that “speeding… is a factor in about 40 percent of road deaths.” Data from the NHTSA puts the number at 30%.

Speeding also increases the *severity* of crashes; “in a 60 km/h speed limit area, the risk of involvement in a casualty crash doubles with each 5 km/h increase in travelling speed above 60 km/h.”

Stop. Think about that for a second. I'll convert it to the Imperial system for my fellow Americans: "*in a [37.3 mph] speed limit area, the risk of involvement in a casualty crash **doubles** with each [3.1 mph] increase in travelling speed above [37.3 mph]."* Remember that next time you drive a 'mere' 5 mph over the limit.

Equally shocking is this paragraph from the Freakonomics blog:

> Kockelman et al. estimated that the difference between a crash on a 55 mph limit road and a crash on a 65 mph one means a **24 percent increase in the chances the accident will be fatal**. Along with the higher incidence of crashes happening in the first place, a difference in limit between 55 and 65 adds up to a 28 percent increase in the overall fatality count.

Driving too slowly can be dangerous too. An NHTSA presentation cites two studies that found a U-shaped relationship between vehicle speed and crash incidence; thus "Crash rates were lowest for drivers traveling near the mean speed, and increased with deviations above and below the mean."

However, driving fast is still far more dangerous than driving slowly. This relationship appears to be exponential, as you can see on the tenth slide of the presentation.

- Solution: Watch this 30 second video for a vivid comparison of head-on crashes at 60 km/hr (37 mph) and 100 km/hr (60 mph). Imagine yourself in the car. Imagine your tearful friends and family.
- Solution: Develop an identity as someone who drives close to the speed limit, by meditating on the commitment, writing a journal entry about it, using Anki, or saying it every day when you wake up in the morning.

Driving conditions

Driving conditions are another source of driving risk.

One factor I discovered was the additional risk from **driving at night.** Nationwide, 49% of fatal crashes happen at night, with **a fatality rate** per mile of travel **about three times as high as daytime hours**. (Source)

- Solution: make an explicit effort to **avoid driving at night.** Use Anki to remember this association.
- Solution: Look at your schedule and see if you can change a recurring night-time drive to the daytime.

Berkeley research on 1.4 million fatal crashes found that "fatal crashes were 14% more likely to happen on the first snowy day of the season compared with subsequent ones." The suggested hypothesis is that people take at least a day to recalibrate their driving behavior in light of new snow.

- Solution: make an explicit effort to **avoid driving on the first snowy day** after a sequence of non-snowy ones. Use Anki to remember this association.

Another valuable factoid: **77%** of weather-related fatalities (and 75% of all crashes!) involve **wet pavement**.

Statistics are available for other weather-related issues, but the data I found wasn't adjusted for the relative frequencies of various weather conditions. That's problematic; it might be that fog, for example, is horrendously dangerous compared to ice or slush, but it's rarer and thus kills fewer people. I'm interested in looking at appropriately adjusted statistics.

Other considerations

- Teen drivers are apparently way worse at not dying in cars than older people. So if you're a teenager, take the outside view and accept that you (not just 'other dumb teenagers') may need to take *particular* care when driving. Relevant information about teen driving is available here.

- Alcohol use appeared so often during my research that I didn't even bother including stats about it. Likewise for wearing a seatbelt.

- Since I'm not in the market for a car, I didn't look into *vehicle choice* as a way to decrease personal existential risk. But I do expect this to be relevant to increasing driving safety.

- "The most dangerous month, it turns out, is August, and Saturday the most dangerous day, according to the National Highway Traffic Safety Administration." I couldn't tell whether this was because of *increased amount of driving* or an *increased rate of crashes*.

- This site recommends driving with your hands at 9 and 3 for increased control. The same site claims that "Most highway accidents occur in the left lane" because the other lanes have "more 'escape routes' should a problem suddenly arise that requires you to quickly change lanes", but I found no citation for the claim.

- Bad driver behavior appears to significantly increase the risk of death in an accident, so: don't ride in car with people who drive badly or aggressively. I have a few friends with aggressive driving habits, and I'm planning to either a) tell them to drive more slowly when I'm in the car or b) stop riding in their cars.

Commenters' recommendations

I should note here that I have not personally verified anything posted below. Be sure to look at the original comment and do followup research before depending on these recommendations.

- MartinB recommends taking a driving safety class every few years.

- Dmytry suggests that bicycling may be good training for constantly keeping one's eyes on the road, though others argue that bicycling itself may be significantly more dangerous than driving anyway.

- Various commenters suggested simply avoiding driving whenever possible. Living in a city with good public transportation is recommended.

- David_Gerard recommends driving a bigger car with larger crumple zones (but not an SUV because they roll over). He also recommends avoiding motorcycles altogether and taking advanced driving courses.

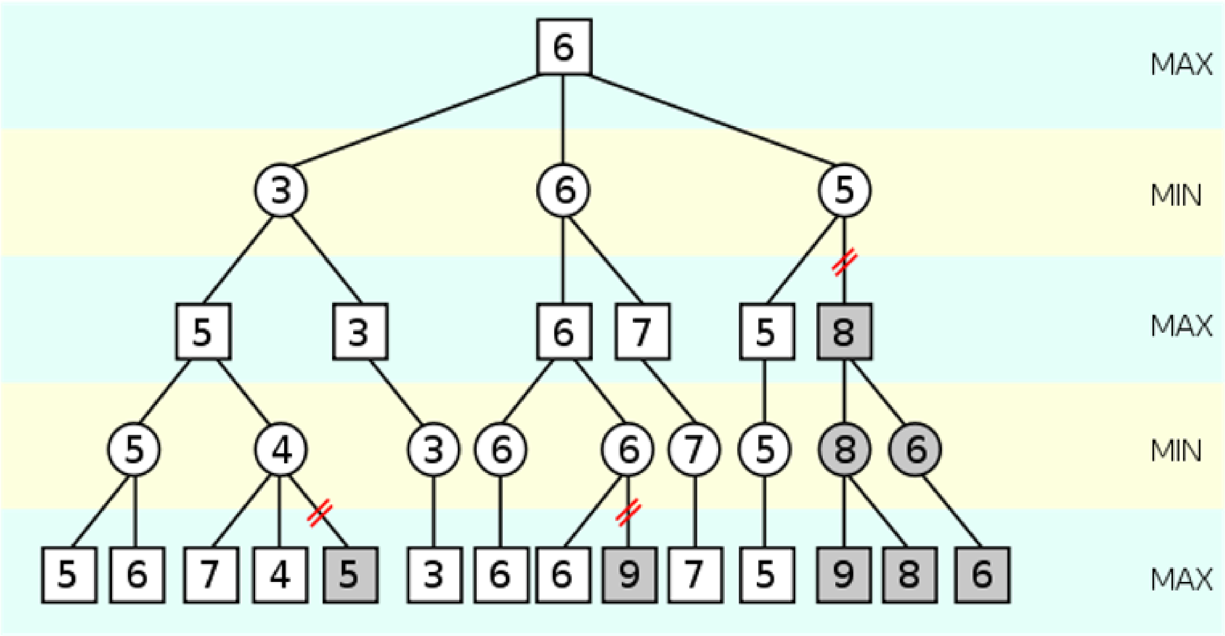
- Craig_Heldreth adds that *everyone* in the car should be buckled up, as even *a single unbuckled passenger* can *collide with and kill other passengers* in a crash. Even cargo as light as a laptop should be secured or put in the trunk.

- JRMayne offers a list of recommendations that merit reading directly. DuncanS also offers a valuable list.

[1] All bolding in the data was added for emphasis by me.

[2] The report notes that "57% of crashes were due solely to driver factors, 27% to combined roadway and driver factors, 6% to combined vehicle and driver factors, 3% solely to roadway factors, 3% to combined roadway, driver, and vehicle factors, 2% solely to vehicle factors and 1% to combined roadway and vehicle factors.”

Decision Theories: A Less Wrong Primer



Summary: *If you've been wondering why people keep going on about decision theory on Less Wrong, I wrote you this post as an answer. I explain what decision theories are, show how Causal Decision Theory works and where it seems to give the wrong answers, introduce (very briefly) some candidates for a more advanced decision theory, and touch on the (possible) connection between decision theory and ethics.*

What is a decision theory?

This is going to sound silly, but a decision theory is an algorithm for making decisions.[0] The inputs are an agent's knowledge of the world, and the agent's goals and values; the output is a particular action (or plan of actions). Actually, in many cases the goals and values are implicit in the algorithm rather than given as input, but it's worth keeping them distinct in theory.

For example, we can think of a chess program as a simple decision theory. If you feed it the current state of the board, it returns a move, which advances the implicit goal of winning. The actual details of the decision theory include things like writing out the tree of possible moves and countermoves, and evaluating which possibilities bring it closer to winning.

Another example is an *E. Coli* bacterium. It has two basic options at every moment: it can use its flagella to swim forward in a straight line, or to change directions by randomly tumbling. It can sense whether the concentration of food or toxin is increasing or decreasing over time, and so it executes a simple algorithm that randomly changes direction more often when things are "getting worse". This is enough control for bacteria to rapidly seek out food and flee from toxins, without needing any sort of advanced information processing.

A human being is a much more complicated example which combines some aspects of the two simpler examples; we mentally model consequences in order to make many decisions, and we also follow heuristics that have evolved to work well without explicitly modeling the world.[1] We can't model anything quite like the complicated way that human beings make decisions, but we can study simple decision theories on simple problems; and the results of this analysis were often more effective than the raw intuitions of human beings (who evolved to succeed in small savannah tribes, not negotiate a nuclear arms race). But the standard model used for this analysis, Causal Decision Theory, has a serious drawback of its own, and the suggested replacements are important for a number of things that Less Wrong readers might care about.

What is Causal Decision Theory?

Causal decision theory (CDT to all the cool kids) is a particular class of decision theories with some nice properties. It's straightforward to state, has some nice mathematical features, can be adapted to any utility function, and gives good answers on many problems. We'll describe how it works in a fairly simple but general setup.

Let **X** be an agent who shares a world with some other agents ($\mathbf{Y_1}$ through $\mathbf{Y_n}$). All these agents are going to privately choose actions and then perform them simultaneously, and the actions will have consequences. (For instance, they could be playing a round of Diplomacy.)

We'll assume that **X** has goals and values represented by a utility function: for every consequence **C**, there's a number **U(C)** representing just how much **X** prefers that outcome, and **X** views equal *expected* utilities with indifference: a 50% chance of utility 0 and 50% chance of utility 10 is no better or worse than 100% chance of utility 5, for instance. (If these assumptions sound artificial, remember that we're trying to make this as mathematically simple as we can in order to analyze it. I don't think it's as artificial as it seems, but that's a different topic.)

X wants to maximize its expected utility. If there were no other agents, this would be simple: model the world, estimate how likely each consequence is to happen if it does this action or that, calculate the expected utility of each action, then perform the action that results in the highest expected utility. But if there are other agents around, the outcomes depend on their actions as well as on **X**'s action, and if **X** treats *that* uncertainty like normal uncertainty, then there might be an opportunity for the **Y**s to exploit **X**.

This is a Difficult Problem in general; a full discussion would involve Nash equilibria, but even that doesn't fully settle the matter- there can be more than one equilibrium! Also, **X** *can* sometimes treat another agent as predictable (like a fixed outcome or an ordinary random variable) and get away with it.

CDT is a *class* of decision theories, not a specific decision theory, so it's impossible to specify with full generality how **X** will decide if **X** is a causal decision theorist. But there is one key property that distinguishes CDT from the decision theories we'll talk about later: a CDT agent assumes that **X**'s decision is *independent* from the simultaneous decisions of the **Y**s- that is, **X** could decide one way or another and everyone else's decisions would stay the same.

Therefore, there is at least one case where we can say what a CDT agent will do in a multi-player game: some strategies are dominated by others. For example, if **X** and **Y** are both deciding whether to walk to the zoo, and **X** will be happiest if **X** and **Y** both go, but **X** would still be happier at the zoo than at home even if **Y** doesn't come along, then **X** should go to the zoo regardless of what **Y** does. (Presuming that **X**'s utility function is focused on being happy that afternoon.) This criterion is enough to "solve" many problems for a CDT agent, and in zero-sum two-player games the solution can be shown to be optimal for **X**.

What's the problem with Causal Decision Theory?

There are many simplifications and abstractions involved in CDT, but that assumption of independence turns out to be key. In practice, people put a lot of effort into predicting what other people might decide, sometimes with impressive accuracy, and then base their own decisions on that prediction. This wrecks the independence of decisions, and so it turns out that in a non-zero-sum game, it's possible to "beat" the outcome that CDT gets.

The classical thought experiment in this context is called Newcomb's Problem. **X** meets with a very smart and honest alien, Omega, that has the power to accurately predict what **X** would do in various hypothetical situations. Omega presents **X** with two boxes, a clear one containing $1,000 and an opaque one containing either $1,000,000 or nothing. Omega explains that **X** can either take the opaque box (this is called *one-boxing*) or both boxes (*two-boxing*), but there's a trick: Omega predicted in advance what **X** would do, and put $1,000,000 into the opaque box only if **X** was predicted to one-box. (This is a little devious, so take some time to ponder it if you haven't seen Newcomb's Problem before- or read here for a fuller explanation.)

If **X** is a causal decision theorist, the choice is clear: whatever Omega decided, it decided already, and whether the opaque box is full or empty, **X** is better off taking both. (That is, two-boxing is a dominant strategy over one-boxing.) So **X** two-boxes, and walks away with $1,000 (since Omega easily predicted that this would happen). Meanwhile, **X**'s cousin **Z** (not a CDT) decides to one-box, and finds the box full with $1,000,000. So it certainly seems that one could do better than CDT in this case.

But is this a fair problem? After all, we can always come up with problems that trick the rational agent into making the wrong choice, while a dumber agent lucks into the right one. Having a very powerful predictor around might seem artificial, although the problem might look much the same if Omega had a 90% success rate rather than 100%. One reason that this is a fair problem is that the outcome depends only on what action **X** is simulated to take, not on what process produced the decision.

Besides, we can see the same behavior in another famous game theory problem: the Prisoner's Dilemma. **X** and **Y** are collaborating on a project, but they have different goals for it, and either one has the opportunity to achieve their goal a little better at the cost of significantly impeding their partner's goal. (The options are called *cooperation* and *defection*.) If they both cooperate, they get a utility of +50 each; if **X** cooperates and **Y** defects, then **X** winds up at +10 but **Y** gets +70, and vice versa; but if they both defect, then both wind up at +30 each.[2]

If **X** is a CDT agent, then defecting dominates cooperating as a strategy, so **X** will always defect in the Prisoner's Dilemma (as long as there are no further ramifications;

the Iterated Prisoner's Dilemma can be different, because **X**'s *current* decision can influence **Y**'s *future* decisions). Even if you knowingly pair up **X** with a copy of itself (with a different goal but the same decision theory), it will defect even though it could prove that the two decisions will be identical.

Meanwhile, its cousin **Z** also plays the Prisoner's Dilemma: **Z** cooperates when it's facing an agent that has the same decision theory, and defects otherwise. This is a strictly better outcome than **X** gets. (**Z** isn't optimal, though; I'm just showing that you can find a strict improvement on **X**.)[3]

What decision theories are better than CDT?

I realize this post is pretty long already, but it's way too short to outline the advanced decision theories that have been proposed and developed recently by a number of people (including Eliezer, Gary Drescher, Wei Dai, Vladimir Nesov and Vladimir Slepnev). Instead, I'll list the features that we would want an advanced decision theory to have:

1. The decision theory should be formalizable at least as well as CDT is.
2. The decision theory should give answers that are at least as good as CDT's answers. In particular, it should always get the right answer in 1-player games and find a Nash equilibrium in zero-sum two-player games (when the other player is also able to do so).
3. The decision theory should strictly outperform CDT on the Prisoner's Dilemma- it should elicit mutual cooperation in the Prisoner's Dilemma from some agents that CDT elicits mutual defection from, it shouldn't cooperate when its partner defects, and (arguably) it should defect if its partner would cooperate regardless.
4. The decision theory should one-box on Newcomb's Problem.
5. The decision theory should be reasonably simple, and not include a bunch of ad-hoc rules. We want to solve problems involving prediction of actions in general, not just the special cases.

There are now a couple of candidate decision theories (Timeless Decision Theory, Updateless Decision Theory, and Ambient Decision Theory) which seem to meet these criteria. Interestingly, formalizing any of these tends to deeply involve the mathematics of self-reference (Gödel's Theorem and Löb's Theorem) in order to avoid the infinite regress inherent in simulating an agent that's simulating you.

But for the time being, we can massively oversimplify and outline them. TDT considers your ultimate decision as the cause of both your action and other agents' valid predictions of your action, and tries to pick the decision that works best under that model. ADT uses a kind of diagonalization to predict the effects of different decisions without having the final decision throw off the prediction. And UDT considers the decision that would be the best policy for all possible versions of you to employ, on average.

Why are advanced decision theories important for Less Wrong?

There are a few reasons. Firstly, there are those who think that advanced decision theories are a natural base on which to build AI. One reason for this is something I briefly mentioned: even CDT allows for the idea that **X**'s current decisions can affect

Y's future decisions, and self-modification counts as a decision. If **X** can self-modify, and if **X** expects to deal with situations where an advanced decision theory would out-perform its current self, then **X** will change itself into an advanced decision theory (with some weird caveats: for example, if **X** started out as CDT, its modification will only care about other agents' decisions made after **X** self-modified).

More relevantly to rationalists, the bad choices that CDT makes are often held up as examples of why you shouldn't try to be rational, or why rationalists can't cooperate. But instrumental rationality doesn't need to be synonymous with causal decision theory: if there are other decision theories that do strictly better, we should adopt those rather than CDT! So figuring out advanced decision theories, even if we can't implement them on real-world problems, helps us see that the ideal of rationality isn't going to fall flat on its face.

Finally, advanced decision theory could be relevant to morality. If, as many of us suspect, there's no basis for human morality apart from what goes on in human brains, then why do we feel there's still a distinction between what-we-want and what-is-right? One answer is that if we feed in what-we-want into an advanced decision theory, then just as cooperation emerges in the Prisoner's Dilemma, many kinds of patterns that we take as basic moral rules emerge as the equilibrium behavior. The idea is developed more substantially in Gary Drescher's Good and Real, and (before there was a candidate for an advanced decision theory) in Douglas Hofstadter's concept of superrationality.

It's still at the speculative stage, because it's difficult to work out what interactions between agents with advanced decision theories would look like (in particular, we don't know whether bargaining would end in a fair split or in a Xanatos Gambit Pileup of chicken threats, though we think and hope it's the former). But it's at least a promising approach to the slippery question of what 'right' could actually mean.

And if you want to understand this on a slightly more technical level... well, I've started a sequence.

Next: A Semi-Formal Analysis, Part I (The Problem with Naive Decision Theory)

Notes:

0. Rather confusingly, decision theory is the name for the study of decision theories.

1. Both patterns appear in our conscious reasoning as well as our subconscious thinking- we care about consequences we can directly foresee and also about moral rules that don't seem attached to any particular consequence. However, just as the simple "program" for the bacterium was constructed by evolution, our moral rules are there for evolutionary reasons as well- perhaps even for reasons that have to do with advanced decision theory...

Also, it's worth noting that we're not consciously aware of all of our values and goals, though at least we have a better idea of them than *E.Coli* does. This is a problem for the idea of representing our usual decisions in terms of decision theory, though we can still hope that our approximations are good enough (e.g. that our real values regarding the Cold War roughly corresponded to our estimates of how bad a nuclear war or a Soviet world takeover would be).

2. Eliezer once pointed out that our intuitions on most formulations of the Prisoner's Dilemma are skewed by our notions of fairness, and a more outlandish example might serve better to illustrate how a genuine PD really feels. For an example where people are notorious for not caring about each others' goals, let's consider aesthetics: people who love one form of music often really feel that another popular form is a waste of time. One might feel that if the works of Artist Q suddenly disappeared from the world, it would objectively be a tragedy; while if the same happened to the works of Artist R, then it's no big deal and R's fans should be glad to be freed from that dreck.

We can use this aesthetic intolerance to construct a more genuine Prisoner's Dilemma without inviting aliens or anything like that. Say **X** is a writer and **Y** is an illustrator, and they have very different preferences for how a certain scene should come across, so they've worked out a compromise. Now, both of them could cooperate and get a scene that both are OK with, or **X** could secretly change the dialogue in hopes of getting his idea to come across, or **Y** could draw the scene differently in order to get her idea of the scene across. But if they both "defect" from the compromise, then the scene gets confusing to readers. If both **X** and **Y** prefer their own idea to the compromise, prefer the compromise to the muddle, and prefer the muddle to their partner's idea, then this is a genuine Prisoner's Dilemma.

3. I've avoided mentioning Evidential Decision Theory, the "usual" counterpart to CDT; it's worth noting that EDT one-boxes on Newcomb's Problem but gives the wrong answer on a classical one-player problem (The Smoking Lesion) which the advanced decision theories handle correctly. It's also far less amenable to formalization than the others.

Fallacies as weak Bayesian evidence

Abstract: *Exactly what is fallacious about a claim like ”ghosts exist because no one has proved that they do not”? And why does a claim with the same logical structure, such as ”this drug is safe because we have no evidence that it is not”, seem more plausible? Looking at various fallacies – the argument from ignorance, circular arguments, and the slippery slope argument - we find that they can be analyzed in Bayesian terms, and that people are generally more convinced by arguments which provide greater Bayesian evidence. Arguments which provide only weak evidence, though often evidence nonetheless, are considered fallacies.*

As a Nefarious Scientist, Dr. Zany is often teleconferencing with other Nefarious Scientists. Negotiations about things such as ”when we have taken over the world, who's the lucky bastard who gets to rule over Antarctica” will often turn tense and stressful. Dr. Zany knows that stress makes it harder to evaluate arguments logically. To make things easier, he would like to build a software tool that would monitor the conversations and automatically flag any fallacious claims as such. That way, if he's too stressed out to realize that an argument offered by one of his colleagues is actually wrong, the software will work as backup to warn him.

Unfortunately, it's not easy to define what counts as a fallacy. At first, Dr. Zany tried looking at the logical form of various claims. An early example that he considered was ”ghosts exist because no one has proved that they do not”, which felt clearly wrong, an instance of the argument from ignorance. But when he programmed his software to warn him about sentences like that, it ended up flagging the claim ”this drug is safe, because we have no evidence that it is not”. Hmm. That claim felt somewhat weak, but it didn't feel obviously wrong the way that the ghost argument did. Yet they shared the same structure. What was the difference?

The argument from ignorance

Related posts: Absence of Evidence is Evidence of Absence, But Somebody Would Have Noticed!

One kind of argument from ignorance is based on *negative evidence.* It assumes that if the hypothesis of interest were true, then experiments made to test it would show positive results. If a drug were toxic, tests of toxicity of reveal this. Whether or not this argument is valid depends on whether the tests *would* indeed show positive results, and with what probability.

With some thought and help from AS-01, Dr. Zany identified three intuitions about this kind of reasoning.

1. Prior beliefs influence whether or not the argument is accepted.

> A) I've often drunk alcohol, and never gotten drunk. Therefore alcohol doesn't cause intoxication.
>
> B) I've often taken Acme Flu Medicine, and never gotten any side effects. Therefore Acme Flu Medicine doesn't cause any side effects.

Both of these are examples of the argument from ignorance, and both seem fallacious. But B seems much more compelling than A, since we *know* that alcohol causes intoxication, while we also know that not all kinds of medicine have side effects.

2. The more evidence found that is compatible with the conclusions of these arguments, the more acceptable they seem to be.

> C) Acme Flu Medicine is not toxic because no toxic effects were observed in 50 tests.
>
> D) Acme Flu Medicine is not toxic because no toxic effects were observed in 1 test.

C seems more compelling than D.

3. Negative arguments are acceptable, but they are generally less acceptable than positive arguments.

> E) Acme Flu Medicine is toxic because a toxic effect was observed (positive argument)
>
> F) Acme Flu Medicine is not toxic because no toxic effect was observed (negative argument, the argument from ignorance)

Argument E seems more convincing than argument F, but F is somewhat convincing as well.

"*Aha!*" *Dr. Zany exclaims.* "*These three intuitions share a common origin! They bear the signatures of Bayonet reasoning!*"

"*Bayesian* *reasoning*", *AS-01 politely corrects.*

"*Yes, Bayesian! But, hmm. Exactly how are they Bayesian?*"

> *Note:* To keep this post as accessible as possible, I attempt to explain the underlying math without actually using any math. If you would rather see the math, please see the paper referenced at the end of the post.

As a brief reminder, the essence of Bayes' theorem is that we have different theories about the world, and the extent to which we believe in these theories varies. Each theory also has implications about what you expect to observe in the world (or at least it *should* have such implications). The extent to which an observation makes us update our beliefs depends on how likely our theories say the observation should be. Dr. Zany has a strong belief that his plans will basically always succeed, and this theory says that his plans are very unlikely to fail. Therefore, when they do fail, he should revise his belief in the "I will always succeed" theory down. (So far he hasn't made that update, though.) If this isn't completely intuitive to you, I recommend komponisto's awesome visualization.

Now let's look at each of the above intuitions in terms of Bayes' theorem.

1. Prior beliefs influence whether or not the argument is accepted. This is pretty straightforward -the expression "prior beliefs" is even there in the description of the intuition. Suppose that we hear the argument, "I've often drunk alcohol, and never gotten drunk. Therefore alcohol doesn't cause intoxication". The fact that this person has never gotten drunk from alcohol (or at least claims that he hasn't) *is* evidence for alcohol not causing any intoxication, but we still have a very strong prior belief for alcohol causing intoxication. Updating on this evidence, we find that our beliefs in both the theory "this person is mistaken or lying" and the theory "alcohol doesn't cause intoxication" have become stronger. Due to its higher prior probability, "this person is mistaken or lying" seems more plausible of the two, so we do not consider this a persuasive argument for alcohol not being intoxicating.

2. The more evidence found that is compatible with the conclusions of these arguments, the more acceptable they seem to be. This too is a relatively straightforward consequence of Bayes' theorem. In terms of belief updating, we might encounter 50 pieces of evidence, one at a time, and make 50 small updates. Or we might encounter all of the 50 pieces of evidence at once, and perform one large update. The end result should be the same. More evidence leads to larger updates.

3. Negative arguments are acceptable, but they are generally less acceptable than positive arguments. This one needs a little explaining, and here we need the concepts of sensitivity and specifity. A test for something (say, a disease) is *sensitive* if it *always* gives a positive result when the disease is present, and *specific* if it *only* gives a positive result when the disease is present. There's a trade-off between these two. For instance, an airport metal detector is designed to alert its operators if a person carries dangerous metal items. It is *sensitive*, because nearly any metal item will trigger an alarm - but it is not very *specific*, because even non-dangerous items will trigger an alarm.

A test which is both extremly sensitive and extremly non-specific is not very useful, since it will give more false alarms than true ones. An easy way of creating an extremely sensitive "test for disease" is to simply *always* say that the patient has the disease. This test has 100% sensitivity (it always gives a positive result, so it always gives a positive result when the disease is present, as well), but its specificity is very low - equal to the prevalence rate of the disease. It provides no information, and isn't therefore a test at all.

How is this related to our intuition about negative and positive arguments? In short, our environment is such that like the airport metal detector, negative evidence often has high sensitivity but low specificity. We intuitively expect that a test for toxicity might not always reveal a drug to be toxic, but if it does, then the drug really *is* toxic. A lack of a "toxic" result is what we would expect if the drug weren't toxic, but it's also what we would expect in a lot of cases where the drug *was* toxic. Thus, negative evidence *is* evidence, but it's usually much weaker than positive evidence.

"*So, umm, okay*", *Dr. Zany says, after AS-01 has reminded him of the way Bayes*' *theorem works, and helped him figure out how his intuitions about the fallacies have Bayes-structure.* "*But let's not lose track of what we were doing, which is to say, building a fallacy-detector. How can we use this to say whether a given claim is fallacious?*"

"*What this suggests is that we judge a claim to be a fallacy if it's only weak Bayesian evidence*", *AS-01 replies.* "*A claim like 'an unreliable test of toxicity didn't reveal this drug to be toxic, so it must be safe' is such weak evidence that we consider it fallacious. Also, if we have a very strong prior belief against something, and a claim doesn't shift this prior enough, then we might call it a 'fallacy' to believe in the thing on the basis of that claim. That was the case with the 'I've had alcohol many times and never gotten drunk, so alcohol must not be intoxicating' claim.*"

"*But that's not what I was after at all! In that case I can't program a simple fallacy-detector: I'd have to implement a full-blown artificial intelligence that could understand the conversation, analyze the prior probabilities of various claims, and judge the weight of evidence. And even if I did that, it wouldn't help me figure out what claims were fallacies, because all of my AIs only want to eradicate the color blue from the universe! Hmm. But maybe the appeal from ignorance was a special case, and other fallacies are more accomodating. How about circular claims? Those must surely be fallacious?*"

<u>Circularity</u>

> A. God exists because the Bible says so, and the Bible is the word of God.
>
> B. Electrons exist because we can see 3-cm tracks in a cloud chamber, and 3-cm tracks in cloud chambers are signatures of electrons.

"*Okay, we have two circular claims here*", *AS-01 notes.* "*Their logical structure seems to be the same, but we judge one of them to be a fallacy, while the other seems to be okay.*"

"*I have a bad feeling about this*", *Dr. Zany says.*

The argument for the fallaciousness of the above two claims is that they presume the conclusion in the premises. That is, it is presumed that the Bible is the word of God, but that is only possible if God actually exists. Likewise, if electrons don't exist, then whatever we see in the cloud chamber isn't the signature signs of electrons. Thus, in order to believe the conclusion, we need to already believe it as an implicit premise.

But from a Bayesian perspective, beliefs aren't binary propositions: we can *tentatively* believe in a hypothesis, such as the existence of God or electrons. In addition to this tentative hypothesis, we have sense data about the existence of the Bible and the 3-cm tracks. This data we take as certain. We also have a second tentative belief, the ambiguous interpretation of this sense data as the word of God or the signature of electrons. The sense data is ambiguous in the sense that it might or might not be the word of God. So we have three components in our inference: the evidence (Bible, 3-cm tracks), the ambiguous interpretation (the Bible is the word of God, the 3-cm tracks are signatures of electrons), and the hypothesis (God exists, electrons exist).

We can conjecture a causal connection between these three components. Let's suppose that God exists (the hypothesis). This then causes the Bible as his word (ambiguous interpretation), which in turn gives rise to the actual document in front of us (sense data). Likewise, if electrons exist (hypothesis), then this can give rise to the predicted signature effects (ambiguous interpretation), which become manifest as what we actually see in the cloud chamber (sense data).

The "circular" claim reverses the direction of the inference. We have sense data, which we would expect to see if the ambiguous interpretation was correct, and we would expect the interpretation to be correct if the hypothesis were true. Therefore it's more likely that the hypothesis is true. Is this allowed? Yes! Take for example the inference "if there are dark clouds in the sky, then it will rain, in which case the grass will be wet". The reverse inference, "the grass is wet, therefore it has rained, therefore there have been dark clouds in the sky" is valid. However, the inference "the grass is wet, therefore the sprinkler has been on, thefore there is a sprinkler near this grass" may *also* be a valid inference. The grass being wet is evidence for *both* the presence of dark clouds *and* for a sprinkler having been on. Which hypothesis do we judge to be more likely? That depends on our prior beliefs about the hypotheses, as well as the strengths of the causal links (e.g. "if there are dark clouds, how likely is it that it rains?", and vice versa).

Thus, the "circular" arguments given above are actually valid Bayesian inferences. But there is a reason that we consider A to be a fallacy, while B sounds valid. Since the intepretation (the Bible is the word of God, 3-cm tracks are signatures of electrons) logically requires the hypothesis, the probability of the interpretation cannot be higher than the probability of the hypothesis. If we assign the existence of God a very low prior belief, then we must also assign a very low prior belief to the interpretation of the Bible as the word of God. In that case, seeing the Bible will not do much to elevate our belief in the claim that God exists, if there are more likely hypotheses to be found.

"*So you're saying that circular reasoning, too, is something that we consider fallacious if our prior belief in the hypothesis is low enough? And recognizing these kinds of fallacies is AI-complete, too?*" *Dr. Zany asks.*

"*Yup!*", *AS-01 replies cheerfully, glad that for once, Dr. Zany gets it without a need to explain things fifteen times.*

"*Damn it. But... what about slippery slope arguments? Dr. Cagliostro claims that if we let minor supervillains stake claims for territory, then we would end up letting henchmen stake claims for territory as well, and eventually we'd give the right to people who didn't even participate in our plans! Surely that must be a fallacy?*"

Slippery slope

Slippery slope arguments are often treated as fallacies, but they might not be. There are cases where the stipulated "slope" *is* what would actually (or likely) happen. For instance, take a claim saying "if we allow microbes to be patented, then that will lead to higher life-forms being patented as well":

> There are cases in law, for example, in which a legal precedent has historically facilitated subsequent legal change. Lode (1999, pp. 511–512) cites the example originally identified by Kimbrell (1993) whereby there is good reason to believe that the issuing of a patent on a transgenic mouse by the U.S. Patent and Trademark Office in the year 1988 is the result of a slippery slope set in motion with the U.S. Supreme court’s decision Diamond v. Chakrabarty. This latter decision allowed a patent for an oil-eating microbe, and the subsequent granting of a patent for the mouse would have been unthinkable without the chain started by it. (Hahn & Oaksford, 2007)

So again, our prior beliefs, here ones about the plausibility of the slope, influence whether or not the argument is accepted. But there is also another component that was missing from the previous fallacies. Because slippery slope arguments are about actions, not just beliefs, the principle of expected utility becomes relevant. A slippery slope argument will be stronger (relative to its alternative) if it invokes a more undesirable potential consequence, if that consequence is more probable, and if the expected utility of the alternatives is smaller.

For instance, suppose for the sake of argument that both increased heroin consumption and increased reggae music consumption are equally likely consequences of cannabis legalization:

A. Legalizing cannabis will lead to an increase in heroin consumption.

> B. Legalizing cannabis will lead to an increase in listening to reggae music.

Yet A would feel like a stronger argument against the legalization of cannabis than argument B, since increased heroin consumption feels like it would have lower utility. On the other hand, if the outcome is shared, then the stronger argument seems to be the one where the causal link seems more probable:

> C. Legalizing Internet access would lead to an increase in the amount of World of Warcraft addicts.
>
> D. Legalizing video rental stores would lead to an increase in the amount of World of Warcraft addicts.

"*Gah. So a strong slippery slope argument is one where both the utility of the outcome,* ***and*** *the outcome's probability is high? So the AI would not only need to evaluate probabilities, but expected utilities as well?*"

"*That's right!*"

"*Screw it, this isn't going anywhere. And here I thought that this would be a productive day.*"

"*They can't all be, but we tried our best. Would you like a tuna sandwich as consolation?*"

"*Yes, please.*"

Because this post is already unreasonably long, the above discussion only covers the *theoretical* reasons for thinking about fallacies as weak or strong Bayesian arguments. For math, experimental studies, and two other subtypes of the argument from ignorance (besides negative evidence), see:

> Hahn, U. & Oaksford, M. (2007) The Rationality of Informal Argumentation: A Bayesian Approach to Reasoning Fallacies. *Psychological Review,* vol. 114, no. 3, 704-732.

I Was Not Almost Wrong But I Was Almost Right: Close-Call Counterfactuals and Bias

Abstract: "*Close-call counterfactuals*", *claims of what could have almost happened but didn't, can be used to either defend a belief or to attack it. People have a tendency to reject counterfactuals as improbable when those counterfactuals threaten a belief (the "I was not almost wrong" defense), but to embrace counterfactuals that support a belief (the "I was almost right" defense). This behavior is the strongest in people who score high on a test for need for closure and simplicity. Exploring counterfactual worlds can be used to reduce overconfidence, but it can also lead to logically incoherent answers, especially in people who score low on a test for need for closure and simplicity.*

”I was not almost wrong”

Dr. Zany, the Nefarious Scientist, has a theory which he intends to use to achieve his goal of world domination. ”As you know, I have long been a student of human nature”, he tells his assistant, AS-01. (Dr. Zany has always wanted to have an intelligent robot as his assistant. Unfortunately, for some reason all the robots he has built have only been interested in eradicating the color blue from the universe. And blue is his favorite color. So for now, he has resorted to just hiring a human assistant and referring to her with a robot-like name.)

”During my studies, I have discovered the following. Whenever my archnemesis, Captain Anvil, shows up at a scene, the media will very quickly show up to make a report about it, and they prefer to send the report live. While this is going on, the whole city – including the police forces! - will be captivated by the report about Captain Anvil, and neglect to pay attention to anything else. This happened once, and a bank was robbed on the other side of the city while nobody was paying any attention. Thus, I know how to commit the perfect crime – I simply need to create a diversion that attracts Captain Anvil, and then nobody will notice me. History tells us that this is the inevitable outcome of Captain Anvil showing up!”

But to Dr. Zany's annoyance, AS-01 is always doubting him. Dr. Zany has often considered turning her into a brain-in-a-vat as punishment, but she makes the best tuna sandwiches Dr. Zany has ever tasted. He's forced to tolerate her impundence, or he'll lose that culinary pleasure.

”But Dr. Zany”, AS-01 says. ”Suppose that some TV reporter had happened to be on her way to where Captain Anvil was, and on her route she saw the bank robbery. Then part of the media attention would have been diverted, and the police would have heard about the robbery. That might happen to you, too!”

Dr. Zany's favorite belief is now being threatened. It might not be inevitable that Captain Anvil showing up will actually let criminals elsewhere act unhindered! AS-01 has presented a plausible-sounding counterfactual, ”if a TV reporter had seen the robbery, then the city's attention had been diverted to the other crime scene”. Although the historical record does not show that Dr. Zany's theory would have been *wrong*, the counterfactual suggests that he might be *almost wrong*.

There are now three tactics that Dr. Zany can use to defend his belief (warrantedly or not):

1. Challenge the mutability of the antecedent. Since AS-01's counterfactual is of the form ”if A, then B”, Dr. Zany could question the plausibility of A.

> ”Baloney!” exclaims Dr. Zany. ”No TV reporter could ever have wandered past, let alone seen the robbery!”

That seems a little hard to believe, however.

2. Challenge the causal principles linking the antecedent to the consequent. Dr. Zany is not logically required to accept the ”then” in ”if A, then B”. There are always unstated background assumptions that he can question.

> ”Humbug!” shouts Dr. Zany. ”Yes, a reporter could have seen the robbery and alerted the media, but given the choice of covering such a minor incident and continuing to report on Captain Anvil, they would not have cared about the bank robbery!”

3. Concede the counterfactual, but insist that it does not matter for the overall theory.

> ”Inconceivable!” yelps Dr. Zany. ”Even if the city's attention would have been diverted to the robbery, the robbers would have escaped by then! So Captain Anvil's presence would have allowed them to succeed regardless!”

Empirical work suggests that it's not only Dr. Zany who wants to stick to his beliefs. Let us for a moment turn our attention away from supervillains, and look at professional historians and analysts of world politics. In order to make sense of something as complicated as world history, experts resort to various simplifying strategies. For instance, one explanatory schema is called *neorealist balancing*. Neorealist balancing claims that ”when one state threatens to become too powerful, other states coalesce against it, thereby preserving the balance of power”. Among other things, it implies that Hitler's failure was predetermined by a fundemental law of world politics.

Tetlock (1998, 1999, 2001) surveyed a number of experts on history and international affairs. He surveyed the experts on their commitment to such theories, and then posed them counterfactuals that conflicted with some of those theories. For instance, counterfactuals that conflicted with neorealist balancing were "If Goering had continued to concentrate Luftwaffe attacks on British airbases and radar stations, Germany would have won the Battle of Britain" and "If the German military had played more effectively on the widespread resentment of local populations toward the Stalinist regime, the Soviet Union would have collapsed". The experts were then asked to indicate the extent to which they agreed with the antecedent, the causal link, and the claim that the counterfactual being true would have substantially changed world history.

As might have been expected, experts who subscribed to a certain theory were skeptical about counterfactuals threatening the theory, and employed all three defenses more than experts who were less committed. Denying the possibility of the antecedent was done the least frequently, while questioning the overall impact of the consequence was the most common defense.

By itself, this might not be a sign of bias – the experts might have been skeptical of a counterfactual because they had an irrational commitment to theory, but they might also have acquired a rational commitment to the theory because they were skeptical of counterfactuals challenging it. Maybe neorealist balancing is true, and the experts subscribing to it are right to defend it. What's more telling is that Tetlock also measured each expert's need for closure. It turned out that if an expert had – like Dr. Zany – had a high need for closure, then they were also more likely to employ defenses questioning the validity of a counterfactual.

> Theoretically, high need-for-closure individuals are characterized by two tendencies: urgency which inclines them to 'seize' quickly on readily available explanations and to dismiss alternatives and permanence which inclines them to 'freeze' on these explanations and persist with them even in the face of formidable counterevidence. In the current context, high need-for-closure individuals were hypothesized to prefer simple explanations that portray the past as inevitable, to defend these explanations tenaciously when confronted by dissonant close-call counterfactuals that imply events could have unfolded otherwise, to express confidence in conditional forecasts that extend these explanations into the future, and to defend disconfirmed forecasts from refutation by invoking second-order counterfactuals that imply that the predicted events almost happened. (Tetlock, 1998)

If two people draw different conclusions from the same information, then at least one of them is wrong. Tetlock is careful to note that the data doesn't reveal whether it's the people with a high or a low need for closure who are closer to the truth, but we probably presume that at least some of them were being exceedingly defensive.

This gives us reason to be worried. If some past occurrance seems to fit perfectly into our pet theory, have we considered the case that we might be almost wrong? And if we have, are we exhibiting an excess need for closure by rushing to its defense, or are we being excessively flexible by unnecessarily admitting that something might have gone differently? We should only admit to being almost wrong if we really were almost wrong, after all. Is the cognitive style we happen to have the one that's the most correlated with getting the right answers?

”I was almost right.”

Having defended his theory against AS-01's criticism, Dr. Zany puts the theory into use by starting a fire in a tar factory, diverting Captain Anvil. While the media is preoccupied with reporting the story, Dr. Zany tries to steal the bridge connecting Example City to the continent. Unfortunately, a City Police patrol boat happens to see this, alerting the police forces (as well as Captain Anvil) to the site. Dr. Zany is forced to withdraw.

”Damn that unanticipated patrol boat!”, Dr. Zany swears. ”If only it had not appeared, my plan would have worked perfectly!” AS-01 wisely says nothing, and avoids being turned into a brain-in-a-vat.

Tetlock (1998, 1999) surveyed a number of experts and asked them to make predictions about world politics. Afterwards, when it was clear whether or not the predictions had turned out to be true, he surveyed them again. It turned out that like Dr. Zany, most of the mistaken experts had not seriously updated their beliefs:

> Not surprisingly, experts who got it right credited their accuracy to their sound reading of the 'basic forces' at play in the situation. Across issue domains they assigned average ratings between 6.5 and 7.6 on a 9-point scale where 9 indicates maximum confidence. Perhaps more surprisingly, experts who got it wrong were almost as likely to believe that their reading of the political situation was fundamentally sound. They assigned average ratings from 6.3 to 7.1, across domain (Tetlock, 1998)

Many of the experts defended their reading of the situation by saying that they were ”almost right”. For instance, experts who predicted in 1988 that the Communist Party of the Soviet Union would grow increasingly authortarian during the next five years were prone to claiming that the hardliner coup of 1991 had almost succeeded, and if that had happened, their prediction would have become true. Similarly, observers of South Africa who in 1988-1989 expected white minority rule to continue or to become increasingly oppressive were likely to believe that were it not for two exceptional individuals – de Klerk and Mandela - in key leadership roles, South Africa could easily have gone the other way.

In total, Tetlock (1999) identified five logically defensible strategies for defending one's forecasts, all of which were employed by at least some of the experts. Again, it was the experts who scored the highest on a need for closure who tended to employ such defenses the most:

1. The antecedent (the A in the ”if A, then B”) was never adequately satisfied. Experts might insist ”if we had properly implemented deterrence or reassurance, we could have averter war” or ”if real shock therapy had been practiced, we could have averted the nasty bout of hyperinflation”.
2. Although the specified antecedent was satisfied, something unexpected happened, severing the normal link of cause and effect. Experts might declare that rapid privatization in state industries would have led to the predicted surge in economic growth, but only if the government had pursued prudent monetary policies.
3. Although the predicted outcome did not occur, it ”almost occurred” and would have, if not for some inherently unpredictable outside shock.
4. Although the predicted outcome has not yet occurred, it eventually will and we just need to be more patient (hardline communists may yet prevail in Moscow, the EU might still fall apart).
5. Although the relevant conditions were satisfied and the predicted outcome never came close to occurring and never will, this should not be held against the framework that inspired the forecast. Forecasts are inherently unreliable and politics is hard to predict: just because the framework failed once didn't mean that it's wrong.

Again, Tetlock is careful to note that although it's tempting to dismiss all such maneuvering as ”transparently defensive post hocery”, it would be wrong to automatically interpret it as bias. Each defense is a potentially valid objection, and might have been the right one to make, in some cases.

But there are also signs of bias. Tetlock (1999) makes a number of observations from his data, noting – among other things – that the stronger the original confidence in a claim, the more likely an expert is to employ various defenses. That would suggest that big threats to an expert's claims of expertise activate many defenses. He also notes that the experts who'd made failed predictions and employed strong defenses tended not to update their confidence, while the experts who'd made failed predictions but didn't employ strong defenses did update.

Again, some of the experts were probably right to defend themselves, but some of them were probably biased and only trying to protect their reputations. We should ourselves be alert when we catch ourselves using one of those techniques to defend our predictions.

Exploring counter-factual worlds: a possible debiasing technique.

”Although my plan failed this time, I was almost right! The next time, I'll be prepared for any patrol boats!”, Dr. Zany mutters to himself, back in the safety of his laboratory.

”Yes, it was an unlikely coincidence indeed”, AS-01 agrees. ”Say, I know that such coincidences are terribly unlikely, but I started wondering – what other coincidence might have caused your plan to fail? Are there any others that we should take into account before the next try?”

”Hmm....”, Dr. Zany responds, thoughtfully.

Tetlock & Lebow (2001) found that experts became less convinced of the inevitability of a scenario when they were explicitly instructed to consider various events that might have led to a different outcome. In two studies, experts were told to consider the Cuban Missile Crisis and, for each day of the crisis, estimate the subjective probability that the crisis would end either peacefully or violently. When experts were told to consider various provided counterfactuals suggesting a different outcome, they thought that a violent outcome remained a possibility for longer than the experts who weren't given such counterfactuals to consider. The same happened when the experts weren't given ready-made counterfactuals, but were told to generate alternative scenarios of their own, at an increasingly fine resolution.

> The other group (n = 34) was asked to consider (1) how the set of more violent endings of the Cuban missile crisis could be disaggregated into subsets in which violence remained localized or spread outside the Caribbean, (2) in turn differentiated into subsets in which violence claimed fewer or more than 100 casualties, and (3) for the higher casualty scenario, still more differentiated into a conflict either limited to conventional weaponry or extending to nuclear. (Tetlock & Lebow, 2001)

Again, the experts who generated counterfactual scenarios became less confident of their predictions. The experts with a low need for closure adjusted their opinions considerably more than the ones with a high need for closure.

However, this technique has its dangers as well. More fine-grained scenarios offer an opportunity to tell more detailed stories, and humans give disproportionate weight to detailed stories. Unpacking the various scenarios leads us to giving too much weight for the individual subscenarios. You might remember the example of ”the USA and Soviet Union suspending relations” being considered less probable than ”the Soviet Union invades Poland, and the USA and Soviet Union suspend relations”, even though the second scenario is a subset of the first. People with a low need for closure seem to be especially suspectible to this, while people with a high need for closure tend to produce more logically coherent answers. This might be considered an advantage of the high need for closure – an unwillingness to engage in extended wild goose chases, and thus assign minor scenarios a disproportionately high probability

References

Tetlock, P.E. (1998) Close-Call Counterfactuals and Belief-System Defenses: I Was Not Almost Wrong But I Was Almost Right. *Journal of Personality and Social Psychology*, Vol. 75, No. 3, 639-652.
http://faculty.haas.berkeley.edu/tetlock/Vita/Philip%20Tetlock/Phil%20Tetlock/1994-1998/1998%20Close-Call%20Counterfactuals%20and%20Belief-System%20Defenses.pdf

Tetlock, P.E. (1999) Theory-Driven Reasoning About Plausible Pasts and Probable Futures in World Politics: Are We Prisoners of Our Preconceptions? *American Journal of Political Science*, Vol. 43, No. 2, 335-366.
http://www.uky.edu/AS/PoliSci/Peffley/pdf/Tetlock%201999%20AJPS%20Theory-driven%20World%20Politics.pdf

Tetlock, P.E. & Lebow, R.N. (2001) Poking Counterfactual Holes in Covering Laws: Cognitive Styles and Historical Reasoning. *American Political Science Review*, Vol. 95, No. 4.
http://faculty.haas.berkeley.edu/tetlock/vita/philip%20tetlock/phil%20tetlock/1999-2000/2001%20poking%20counterfactual%20holes%20in%20covering%20laws....pdf

The Stable State is Broken

or: *Why Everything Is Terrible, An Overview.*[1]

It sounds like a theory which explains too much. But it's not a theory, hardly even an explanation, more a pattern that manifests itself once you start trying to seriously answer rhetorical questions about the state of the world. From many perspectives, it's obvious to the point of being mundane, practically tautological, but sometimes such obvious facts are worth pointing out regardless.

The idea is this: **The subset of participants which rises to prominence in any area does so because its members have traits helpful to *becoming prominent*, not necessarily because they have traits which are desirable.** Thus, without ongoing and concerted effort, a great many arenas end up dominated by players employing strategies which are bad for everyone.

This comes up again and again:

- Why does science (or rather, the publisher-based model thereof) so frequently produce results which are laughably wrong? Because those journals which don't publish retractions or reproductions will more frequently be the first to publish revolutionary results, and so become more widely read and widely cited. Journals don't attract authors by being as accurate as possible; they win by looking important.
- Why do cigarette companies target kids and teens whenever they think they can get away with it, and breed tobacco for maximized nicotine? Because those companies which do will turn more profit and thus last longer and grow faster than those that don't, and so have more resources to devote to proliferating. Companies don't expand by playing fair; they win by making and keeping customers.
- Why is the Make-A-Wish Foundation sitting on more donations than it knows what to do with when the Against Malaria Foundation could have used that money to save literally tens or hundreds of thousands of lives per year? Because knowing how to elicit donations is a skill almost completely unrelated to knowing how to spend donations, and because American children with cancer make for better advertising than African children with malaria. Charities don't get donations by making the best possible use of their money; they win by advertising effectively towards potential donors. (cf. Efficient Charity)
- Why do governments inevitably end up run by career lawyers and politicians instead of scientists and economists[2]? Because polarizing rhetoric and political connections *look better* than a nuanced, accurate understanding of the issues. There is only finite time for training and practice, and eventually a choice must be made between training in looking good and training in *being* good. People don't get elected or appointed by being good Bayesians; they win by being popular.
- Why do the big media channels seem to be more concerned with celebrities than science, and spend more time talking about individual murders than they do entire genocides? Because those channels talking about Laci Peterson seem

more personal and are thus more watched than those talking about some religious sect in China. Television programming isn't determined by what's important; what wins is what's watched.

- Why is the sex ratio in animals almost always nearly 1:1, when a population with one male for every five females could grow faster and adapt to problems more readily? Because in such a population, or in any population with a sufficiently large gender imbalance, a gene causing a woman to only have male children will be vastly overrepresented in the grandchild generation relative to the rest of the population, and so shift the balance closer to 50/50. Genes don't proliferate by being good for the species; they win by being good for themselves. (cf. Evolutionarily stable strategy, evolutionary game theory.)
- Why do most big businesses make use of sweatshop conditions and shady tax dodges? Because the businesses which do so will outperform the businesses which don't. Corporations don't grow by being nice; they win by being profitable.
- Why do so many apparently intelligent people spend hours per day idly browsing the likes of Reddit, Hacker News, or TVTropes (or indeed LW), when a similar dedication to active self-improvement could have made them a master of a field inside of a decade? (Using for back-of-the-envelope's sake the supposition that 10,000 hours of practice are required for mastery of some specific art, we find that three hours per day for ten years is approximately 1.1 masteries.) Because which activities become habitual is determined by their immediate dopamine release, and for intelligent people the act of (say) reading about strategies for becoming an effective entrepreneur makes for more instant dopamine than does the painful daily grind involved in *actually* becoming an entrepreneur. Activities don't become part of daily life by being useful; they win by tricking your brain into making them feel good.

It's extremely important to remember that none of this requires active malice, not even foresight or awareness of the strategy utilized. If someone or something *happens* upon a strategy like those described above, it will outperform its peers and become more widespread. This requires no conspiracy, no evil forces at work in the world, not even any individual shifting in their personal stance; these are just the stable strategies towards which the set of *surviving* players eventually converges.

The next question: What can we do about it?

[1] I have distinct recollections of having read an article much like the one I've written here at some point in the past. However, I can't find said article, so at the least we can let this article serve as a refresher or another viewpoint on the matter. (ETA: evgenit and gwern both point out that the article I'm thinking of is Scott Aaronson's Malthusianisms. [Aaronson refers to these states as Nash equilibria, which is not strictly correct; there's no underlying assumption about the rationality of the players here. You don't need *intelligent* participants for selection to operate. This is more a quibble over terminology than anything.])

[2] Until recently, China stood as a notable exception; now it appears that the next generations of leaders will have built their entire careers on shilling the party line.

[3] Tangentially related reading material: Bruce Schneier's new book.

[4] No, neither footnote 3 nor this footnote actually have corresponding backreferences.

[5] I wasn't entire sure which section of the site this would be best suited for. Hopefully this is appropriate. ETA: Also, as this is my first submission outside of comment threads, any feedback is highly appreciated.

Biased Pandemic

Recently, Portland Lesswrong played a game that was a perfect trifecta of: difficult mental exercise; fun; and an opportunity to learn about biases and recognize them in yourself and others. We're still perfecting it, and we'd welcome feedback, especially from people who try it.

The Short Version

The game is a combination of Pandemic, a cooperative board game that is cognitively demanding, and the idea of roleplaying cognitive biases. Our favorite way of playing it (so far), everyone selects a bias at random, and then attempts to exaggerate that bias in their arguments and decisions during the game. Everyone attempts to identify the biases in the other players, and, when a bias is guessed, the guessed player selects a new bias and begins again.

The Pieces

First, Pandemic. Pandemic is a cooperative game with a win condition and three lose conditions that are separate. It provides each player a list of available actions and then allows them 4 moves in which to mix and match those actions. Because of the combined win and lose conditions, players are constantly forced to decide between tactical and strategic objectives, with a strong emphasis on making it easy to choose tactically good moves at the expense of failing to win the game and thus losing by taking too long. Pandemic is a fun game, and it's a great game for people looking to stretch their brains. Obviously you want to be familiar with, preferably experienced at Pandemic before attempting Biased Pandemic. We did have one inexperienced player at Pandemic (out of 4) and that seemed to work okay, though it may have been harder for him than the rest of us.

Enter the biases. Each player selects a bias. We printed out the biases listed here and used them to select our biases. One of our players just made a TOC for selecting biases. There happen to be 104 biases listed in that document, so a deck of cards combined with a coin flip allowed for bias selection. A computer's random number generator, dice, or any other random method should suffice. Some biases may seem unplayable to some players -- certainly, the monetary biases seem unplayable to most of us -- but other players may find a way to play it, so we've refused to cross off biases and just allowed players to re-roll if they get a bias they're sure they can't play.

Examples

This can be a little difficult to wrap your brain around, so let me give a couple of examples. One player, playing the Negativity Bias, went around the board treating cities which had outbroken earlier in the game and ignoring other issues. Another player with Hyperbolic Discounting went further: he treated cities, any city near him, while carrying 5 red city cards in his hand and pointing out, in response to entreaties to cure red, that red wasn't much of an issue right now. A player with Reactance had the winning yellow card and simply refused to be told to go somewhere to give it to the player with the other four. He even went so far as to refuse a half a dozen offers of an airlift so he could give up that card. A player with Hindsight Bias claimed that he had predicted that the player with 1 red card would get two more on his next draw, and was upset that he'd let the other players argue him otherwise. A player with The Ultimate Attribution Error suggested that if we weren't doing well because no rationalist could ever win this game because we were terrible at it. A player with the Authority Bias attempted to suggest that we should do things because it's what Eliezer would want us to do. A player with Illusion of Control declared that his next draw, he simply would not draw an epidemic. There were many others.

Recommended Rules Of Play

We played it somewhat haphazardly the first game, but at the end we agreed on a structure for the next game that we think is better. In our next game we plan to have the order of play go like this: during each player's turn, all players can discuss what the player should do for a timed interval, perhaps 1-2 minutes. The player then declares their intended move. Now each other player gets an opportunity to make a single bias guess. If a guess is correct, the player stops playing the bias, and begins the round again. At the end of their turn, if their bias was guessed, they select a new bias. We considered a bias to be guessed correctly if the player guessing fully described the bias, not just the biased behavior. Bias names, however, were not required.

Notes

One way that you can not do well is by falling into the trap of making the same biased statements repeatedly. After a few rounds of this, the biased statements were pretty obvious. The guesses are an indicator of what the other players are seeing, and we went out of our way to look for ways to respond to the guesses by playing up the aspects of the bias that the other players weren't seeing. For example, a lot of the different biases look like simple overconfidence. One player was playing The Illusion of Control in such a way that the rest of us thought he was overconfident. His response was to start declaring that he simply wasn't going to draw an epidemic card, and when he drew one, he declared that it was my fault for making him draw the card. This was obviously not simply overconfidence.

Before playing, you should figure out how familiar you are with the biases. Players who are incredibly familiar with all of the biases may want to play a game where everyone plays as subtly as possible and your goal is to prevent other people from noticing your bias. For us, our goal was to learn the biases better and identify them in other people, so we tried to ham it up and play them as obviously as possible at first. It was incredibly difficult to specifically identify the biased thinking behind obviously biased statements, even with that, so I'd suggest at least trying it with obviousness first.

One of the most difficult things to remember is that your goal is not to win the Pandemic game. Sure, that's nice, but your real goal is to familiarize yourself with biases, and to have fun roieplaying and identifying biases. Losing Pandemic, especially because the players are following their biased thinking, is a totally acceptable outcome. We won, and do not credit our thinking for it.

We're looking forward to trying the game again, and maybe you'll have suggestions for improving it.

Common mistakes people make when thinking about decision theory

From my experience reading and talking about decision theory on LW, it seems that many of the unproductive comments in these discussions can be attributed to a handful of common mistakes.

Mistake #1: Arguing about assumptions

The main reason why I took so long to understand Newcomb's Problem and Counterfactual Mugging was my insistence on denying the assumptions behind these puzzles. I could have saved months if I'd just said to myself, okay, is this direction of inquiry interesting when taken on its own terms?

Many assumptions seemed to be divorced from real life at first. People dismissed the study of electromagnetism as an impractical toy, and considered number theory hopelessly abstract until cryptography arrived. The only way to make intellectual progress (either individually or as a group) is to explore the implications of interesting assumptions wherever they might lead. Unfortunately people love to argue about assumptions instead of getting anything done, though they can't really judge before exploring the implications in detail.

Several smart people on LW are repeating my exact mistake about Newcomb's Problem now, and others find ways to commit the same mistake when looking at our newer ideas. It's so frustrating and uninteresting to read yet another comment saying my assumptions look unintuitive or unphysical or irrelevant to FAI or whatever. I'm not against criticism, but somehow such comments never blossom into interesting conversations, and that's reason enough to caution you against the way of thinking that causes them.

Mistake #2: Stopping when your idea seems good enough

There's a handful of ideas that decision theory newbies rediscover again and again, like pointing out indexical uncertainty as the solution to Newcomb's problem, or adding randomness to models of UDT to eliminate spurious proofs. These ideas don't work and don't lead anywhere interesting, but that's hard to notice when you just had the flash of insight and want to share it with the world.

A good strategy in such situations is to always push a little bit *past* the point where you have everything figured out. Take one extra step and ask yourself: "Can I make this idea precise?" What are the first few implications? What are the obvious extensions? If your result seems to contradict what's already known, work through some of the contradictions yourself. If you don't find any mistakes in your idea, you will surely find new formal things to say about your idea, which always helps.

Mistake #2A: Stopping when your idea actually is good enough

I didn't want to name any names in this post because my status on LW puts me in a kinda position of power, but there's a name I can name with a clear conscience. In 2009, Eliezer wrote:

> Formally you'd use a Godelian diagonal to write (...)

Of course that's not a newbie mistake at all, but an awesome and fruitful idea! As it happens, writing out that Godelian diagonal immediately leads to all sorts of puzzling questions like "but what does it actually do? and how do we prove it?", and eventually to all the decision theory research we're doing now. Knowing Eliezer's intelligence, he probably could have preempted most of our results. Instead he just declared the problem solved. Maybe he thought he was already at 0.95 formality and that going to 1.0 would be a trivial step? I don't want to insinuate here, but IMO he made a mistake.

Since this mistake is indistinguishable from the last, the remedy for it is the same: "Can I make this idea precise?" Whenever you stake out a small area of knowledge and make it amenable to mathematical thinking, you're likely to find new math that has lasting value. When you stop because your not-quite-formal idea seems already good enough, you squander that opportunity.

...

If this post has convinced you to stop making these common mistakes, be warned that it won't necessarily make you happier. As you learn to see more clearly, the first thing you'll see will be a locked door with a sign saying "Research is hard". Though it's not very scary or heroic, mostly you just stand there feeling stupid about yourself :-)

How to Fix Science

Like The Cognitive Science of Rationality, this is a post for beginners. Send the link to your friends!

Science is broken. We know why, and we know how to fix it. What we lack is the will to change things.

In 2005, several analyses suggested that most published results in medicine are false. A 2008 review showed that perhaps 80% of academic journal articles mistake "statistical significance" for "significance" in the colloquial meaning of the word, an elementary error every introductory statistics textbook warns against. This year, a detailed investigation showed that half of published neuroscience papers contain one particular simple statistical mistake.

Also this year, a respected senior psychologist published in a leading journal a study claiming to show evidence of precognition. The editors explained that the paper was accepted because it was written clearly and followed the usual standards for experimental design and statistical methods.

Science writer Jonah Lehrer asks: "Is there something wrong with the scientific method?"

Yes, there is.

This shouldn't be a surprise. What we currently call "science" isn't the *best* method for uncovering nature's secrets; it's just the first set of methods we've collected that *wasn't totally useless* like personal anecdote and authority generally are.

As time passes we learn new things about how to do science better. The Ancient Greeks practiced some science, but few scientists tested hypotheses against mathematical models before Ibn al-Haytham's 11th-century *Book of Optics* (which also contained hints of Occam's razor and positivism). Around the same time, Al-Biruni emphasized the importance of repeated trials for reducing the effect of accidents and errors. Galileo brought mathematics to greater prominence in scientific method, Bacon described eliminative induction, Newton demonstrated the power of consilience (unification), Peirce clarified the roles of deduction, induction, and abduction, and Popper emphasized the importance of falsification. We've also discovered the usefulness of peer review, control groups, blind and double-blind studies, plus a variety of statistical methods, and added these to "the" scientific method.

In many ways, the best science done today is better than ever — but it still has problems, and most science is done poorly. The good news is that we know what these problems are and we know multiple ways to fix them. What we lack is the *will* to change things.

This post won't list all the problems with science, nor will it list all the promising solutions for any of these problems. (Here's one I left out.) Below, I only describe a few of the basics.

Problem 1: Publication bias

When the study claiming to show evidence of precognition was published, psychologist Richard Wiseman set up a registry for advance announcement of new attempts to replicate the study.

Carl Shulman explains:

> A replication registry guards against publication bias, and at least 5 attempts were registered. As far as I can tell, all of the subsequent replications have, unsurprisingly, failed to replicate Bem's results. However, JPSP and the other high-end psychology journals refused to publish the results, citing standing policies of not publishing straight replications.
>
> From the journals' point of view, this (common) policy makes sense: bold new claims will tend to be cited more and raise journal prestige (which depends on citations per article), even though this means most of the 'discoveries' they publish will be false despite their low p-values (high statistical significance). However, this means that overall the journals are giving career incentives for scientists to massage and mine their data for bogus results, but not to challenge bogus results presented by others.

This is an example of publication bias:

> Publication bias is the term for what occurs whenever the research that appears in the published literature is systematically unrepresentative of the population of completed studies. Simply put, when the research that is readily available differs in its results from the results of *all* the research that has been done in an area, readers and reviewers of that research are in danger of drawing the wrong conclusion about what that body of research shows. In some cases this can have dramatic consequences, as when an ineffective or dangerous treatment is falsely viewed as safe and effective. [Rothstein et al. 2005]

Sometimes, publication bias can be more deliberate. The anti-inflammatory drug Rofecoxib (Vioxx) is a famous case. The drug was prescribed to 80 million people, but in it was later revealed that its maker, Merck, had withheld evidence of the drug's risks. Merck was forced to recall the drug, but it had already resulted in 88,000-144,000 cases of serious heart disease.

Example partial solution

One way to combat publication bias is for journals to only accept experiments that were registered in a public database before they began. This allows scientists to see which experiments were conducted but never reported (perhaps due to negative results). Several prominent medical journals (e.g. *The Lancet* and *JAMA*) now operate this way, but this protocol is not as widespread as it could be.

Problem 2: Experimenter bias

Scientists are humans. Humans are affected by cognitive heuristics and biases (or, really, humans just *are* cognitive heuristics and biases), and they respond to incentives that may not align with an optimal pursuit of truth. Thus, we should expect *experimenter bias* in the practice of science.

There are many stages in research during which experimenter bias can occur:

1. in reading-up on the field,
2. in specifying and selecting the study sample,
3. in [performing the experiment],
4. in measuring exposures and outcomes,
5. in analyzing the data,
6. in interpreting the analysis, and
7. in publishing the results. [Sackett 1979]

Common biases have been covered elsewhere on Less Wrong, so I'll let those articles explain how biases work.

Example partial solution

There is some evidence that the skills of rationality (e.g. cognitive override) are teachable. Training scientists to notice and meliorate biases that arise in their thinking may help them to reduce the magnitude and frequency of the thinking errors that may derail truth-seeking attempts during each stage of the scientific process.

Problem 3: Bad statistics

I remember when my statistics professor first taught me the reasoning behind "null hypothesis significance testing" (NHST), the standard technique for evaluating experimental results. NHST uses "p-values," which are statements about the probability of getting some data (e.g. one's experimental results) *given* the hypothesis being tested. I asked my professor, "But don't we want to know the probability of the hypothesis we're testing *given* the data, not the other way around?" The reply was something about how this was the best we could do. (But that's false, as we'll see in a moment.)

Another problem is that NHST computes the probability of getting data as unusual as the data one collected by considering what might be expected if that particular experiment was repeated many, many times. But how do we know anything about these imaginary repetitions? If I want to know something about a particular earthquake, am I supposed to imagine a few dozen repetitions of that earthquake? What does that even *mean*?

I tried to answer these questions on my own, but all my textbooks assumed the soundness of the mistaken NHST framework for scientific practice. It's too bad I didn't have a class with biostatistican Steven Goodman, who says:

> The p-value is almost nothing sensible you can think of. I tell students to give up trying.

The sad part is that the logical errors of NHST are old news, and have been known ever since Ronald Fisher began advocating NHST in the 1920s. By 1960, Fisher had out-advocated his critics, and philosopher William Rozeboom remarked:

> Despite the awesome pre-eminence [NHST] has attained... it is based upon a fundamental misunderstanding of the nature of rational inference, and is seldom if ever appropriate to the aims of scientific research.

There are many more problems with NHST and with "frequentist" statistics in general, but the central one is this: NHST does not follow from the axioms (foundational logical rules) of probability theory. It is a grab-bag of techniques that, depending on how those techniques are applied, can lead to *different* results when analyzing the *same data* — something that should horrify every mathematician.

The inferential method that solves the problems with frequentism — and, more importantly, follows deductively from the axioms of probability theory — is Bayesian inference.

So why aren't *all* scientists using Bayesian inference instead of frequentist inference? Partly, we can blame the vigor of NHST's early advocates. But we can also attribute NHST's success to the simple fact that Bayesian calculations can be *more difficult* than frequentist calculations. Luckily, new software tools like WinBUGS let computers do most of the heavy lifting required for Bayesian inference.

There's also the problem of sheer momentum. Once a practice is enshrined, it's hard to dislodge it, even for good reasons. I took three statistics courses in university and *none* of my textbooks mentioned Bayesian inference. I didn't learn about it until I dropped out of university and studied science and probability theory on my own.

Remember the study about precognition? Not surprisingly, it was done using NHST. A later Bayesian analysis of the data disconfirmed the original startling conclusion.

Example partial solution

This one is obvious: teach students probability theory instead of NHST. Retrain current scientists in Bayesian methods. Make Bayesian software tools easier to use and more widespread.

Conclusion

If I'm right that there is unambiguous low-hanging fruit for improving scientific practice, this suggests that particular departments, universities, or private research institutions can (probabilistically) out-perform their rivals (in terms of actual discoveries, not just publications) given similar resources.

I'll conclude with one particular *specific* hypothesis. If I'm right, then a research group should be able to hire researchers trained in Bayesian reasoning and in catching publication bias and experimenter bias, and have them extract from the existing literature valuable medical truths that the mainstream medical community doesn't yet know about. This prediction, in fact, is about to be tested.

SotW: Check Consequentialism

(The Exercise Prize series of posts is the Center for Applied Rationality asking for help inventing exercises that can teach cognitive skills. The difficulty is coming up with exercises interesting enough, with a high enough hedonic return, that people actually do them and remember them; this often involves standing up and performing actions, or interacting with other people, not just working alone with an exercise booklet and a pencil. We offer prizes of $50 for any suggestion we decide to test, and $500 for any suggestion we decide to adopt. This prize also extends to LW meetup activities and good ideas for verifying that a skill has been acquired. See here for details.)

Exercise Prize: Check Consequentialism

In philosophy, "consequentialism" is the belief that doing the right thing makes the world a better place, i.e., that actions should be chosen on the basis of their probable outcomes. It seems like the mental habit of *checking consequentialism,* asking "What positive future events does this action cause?", would catch numerous cognitive fallacies.

For example, the mental habit of consequentialism would counter the sunk cost fallacy - if a PhD wouldn't really lead to much in the way of desirable job opportunities or a higher income, and the only reason you're still pursuing your PhD is that *otherwise all your previous years of work will have been wasted,* you will find yourself encountering a blank screen at the point where you try to imagine a positive *future* outcome of spending another two years working toward your PhD - you will not be able to state what good future events happen as a result.

Or consider the problem of *living in the should-universe;* if you're thinking, *I'm not going to talk to my boyfriend about X because he* should *know it already,* you might be able to spot this as an instance of should-universe thinking (planning/choosing/acting/feeling as though within / by-comparison-to an image of an ideal perfect universe) by having done exercises specifically to sensitize you to should-ness. *Or,* if you've practiced the *more general* skill of Checking Consequentialism, you might notice a problem on asking "What happens if I talk / don't talk to my boyfriend?" - providing that you're sufficiently adept to constrain your consequentialist visualization to what *actually* happens as opposed to what *should* happen.

Discussion:

The skill of Checking Consequentialism isn't quite as simple as telling people to ask, "What positive result do I get?" By itself, this mental query is probably going to return *any* apparent justification - for example, in the sunk-cost-PhD example, asking "What good thing happens as a result?" will just return, "All my years of work won't have been wasted! That's good!" Any choice people are tempted by seems good for *some* reason, and executing a query about "good reasons" will just return this.

The novel part of Checking Consequentialism is the ability to *discriminate* "consequentialist reasons" from "non-consequentialist reasons" - being able to distinguish that "Because a PhD gets me a 50% higher salary" talks about future positive consequences, while "Because I don't want my years of work to have been wasted" doesn't.

It's possible that asking "At what time does the consequence occur and how long does it last?" would be useful for distinguishing future-consequences from non-future-consequences - if you take a bad-thing like "I don't want my work to have been wasted" and ask "When does it occur, where does it occur, and how long does it last?", you will with luck notice the error.

Learning to draw cause-and-effect directed graphs, a la Judea Pearl and Bayes nets, seems like it might be helpful - at least, Geoff was doing this while trying to teach strategicness and the class seemed to like it.

Sometimes non-consequentialist reasons can be rescued as consequentialist ones. "You shouldn't kill because it's the wrong thing to do" can be rescued as "Because then a person will transition from 'alive' to 'dead' in the future, and this is a bad event" or "Because the interval between Outcome A and Outcome B includes the interval from Fred alive to Fred dead."

On a five-second level, the skill would have to include:

- Being cued by some problem to try looking at the consequences;
- Either directly having a mental procedure that *only* turns up consequences, like trying to visualize events out into the future, *or*
- First asking 'Why am I doing this?' and then looking at the justifications to check if they're consequentialist, perhaps using techniques like asking 'How long does it last?', 'When does it happen?', or 'Where does it happen?'.
- Expending a *small* amount of effort to see *if* a non-consequentialist reason can easily translate into a consequentialist one *in a realistic way.*
- Making the decision whether or not to change your mind.
- If necessary, detaching from the thing you were doing for non-consequentialist reasons.

In practice, it may be obvious that you're making a mistake as soon as you think to check consequences. I have 'living in the should-universe' or 'sunk cost fallacy' cached to the point where as soon as I spot an error of that pattern, it's usually pretty obvious (without further deliberative thought) what the residual reasons are and whether I was doing it wrong.

Pain points & Pluses:

(When generating a candidate kata, almost the first question we ask - directly after the selection of a topic, like 'consequentialism' - is, "What are the pain points? Or pleasure points?" This can be errors you've made yourself and noticed afterward, or even cases where you've noticed someone else doing it wrong, but ideally cases where you use the skill in real life. Since a lot of rationality is in fact about not screwing up, there may not always be pleasure points where the skill is used in a non-error-correcting, strictly positive context; but it's still worth asking each time. We ask this question right at the beginning because it (a) checks to see how often the skill is actually important in real life and (b) provides concrete use-cases to focus discussion of the skill.)

Pain points:

Checking Consequentialism looks like it should be useful for countering:

- Living in the should-universe (taking actions because of the consequences they *ought* to have, rather than the consequences they *probably will* have). E.g., "I'm not going to talk to my girlfriend because she should already know X" or "I'm going to become a theoretical physicist because I ought to enjoy theoretical physics."
- The sunk cost fallacy (choosing to prevent previously expended, non-recoverable resources from *having been wasted in retrospect* - i.e., avoiding the mental pain of reclassifying a *past* investment as a loss - rather than acting for the sake of future considerations). E.g., "If I give up on my PhD, I'll have wasted the last three years."
- Cached thoughts and habits; "But I usually shop at Whole Foods" or "I don't know, I've never tried an electric toothbrush before." (These might have rescuable consequences, but as stated, they aren't talking about future events.)
- Acting-out an emotion - one of the most useful pieces of advice I got from Anna Salamon was to find other ways to act out an emotion than strategic choices. If you're feeling frustrated with a coworker, you might still want to Check Consequentialism on "Buy them dead flowers for their going-away party" even though it seems to express your frustration.
- Indignation / acting-out of morals - "Drugs are bad, so drug use ought to be illegal", where it's much harder to make the case that countries which decriminalized marijuana experienced worse net outcomes. (Though it should be noted that you also have to Use Empiricism to ask the question 'What happened to other countries that decriminalized marijuana?' instead of making up a gloomy consequentialist prediction to express your moral disapproval.)
- Identity - "I'm the sort of person who belongs in academia."
- "Trying to do things" for simply no reason at all, while your brain still generates activities and actions, because nobody ever told you that behaviors ought to have a purpose or that lack of purpose is a warning sign. This habit can be inculcated by schoolwork, wanting to put in 8 hours before going home, etc. E.g. you "try to write an essay", and you know that an essay has paragraphs; so you try to write a bunch of paragraphs but you don't have any functional role in mind for each paragraph. "What is the positive consequence of this paragraph?" might come in handy here.

(This list is not intended to be exhaustive.)

Pleasure points:

- Being able to state and then focus on a positive outcome seems like it should improve motivation, at least in cases where the positive outcome is realistically attainable to a non-frustrating degree and has not yet been subject to hedonic adaptation. E.g., a $600 job may be more motivating if you visualize the $600 laptop you're going to buy with the proceeds.

Also, consequentialism is the foundation of expected utility, which is the foundation of instrumental rationality - this is why we're considering it as an early unit. (This is not directly listed as a "pleasure point" because it is not directly a use-case.)

Constantly asking about consequences seems likely to improve overall strategicness - not just lead to the better of two choices being taken from a fixed decision-set, but also having goals in mind that can generate new perceived choices, i.e., improve the overall degree to which people do things for reasons, as opposed to not doing things or not having reasons. (But this is a hopeful eventual positive consequence of practicing the skill, not a use-case where the skill is directly being applied.)

Teaching & exercises:

This is the part that's being thrown open to Less Wrong generally. Hopefully I've described the skill in enough detail to convey what it *is.* Now, how would you *practice* it? How would you have an audience practice it, hopefully in *activities* carried out with *each other?*

The dumb thing I tried to do previously was to have exercises along the lines of, "Print up a booklet with little snippets of scenarios in them, and ask people to circle non-consequentialist reasoning, then try to either translate it to consequentialist reasons or say that no consequentialist reasons could be found." I didn't do that for this exact session, but if you look at what I did with the sunk cost fallacy, it's the same sort of silly thing I tried to do.

This didn't work very well - maybe the exercises were too easy, or maybe it was that people were doing it alone, or maybe we did something else wrong, but the audience appeared to experience insufficient hedonic return. They were, in lay terms, unenthusiastic.

At this point I should like to pause, and tell a recent and important story. On Saturday I taught an 80-minute unit on Bayes's Rule to an audience of non-Sequence-reading experimental subjects, who were mostly either programmers or in other technical subjects, so I could go through the math fairly fast. Afterward, though, I was worried that they hadn't really learned to apply Bayes's Rule and wished I had a small little pamphlet of practice problems to hand out. I still think this would've been a good idea, but...

On *Wednesday,* I attended Andrew Critch's course at Berkeley, which was roughly mostly-instrumental LW-style cognitive-improvement material aimed at math students; and in this particular session, Critch introduced Bayes's Theorem, not as advanced math, but with the aim of getting them to apply it to life.

Critch demonstrated using what he called the Really Getting Bayes game. He had Nisan (a local LWer) touch an object to the back of Critch's neck, a cellphone as it happened, while Critch faced in the other direction; this was "prior experience". Nisan said that the object was either a cellphone or a pen. Critch gave prior odds of 60% : 40% that the object was a cellphone vs. pen, based on his prior experience. Nisan then asked Critch how likely he thought it was that a cellphone or a pen would be RGB-colored, i.e., colored red, green, or blue. Critch didn't give exact numbers here, but said he thought a cellphone was more likely to be primary-colored, and drew some rectangles on the blackboard to illustrate the likelihood ratio. After being told that the object was in fact primary-colored (the cellphone was metallic blue), Critch gave posterior odds of 75% : 25% in favor of the cellphone, and then turned around to look.

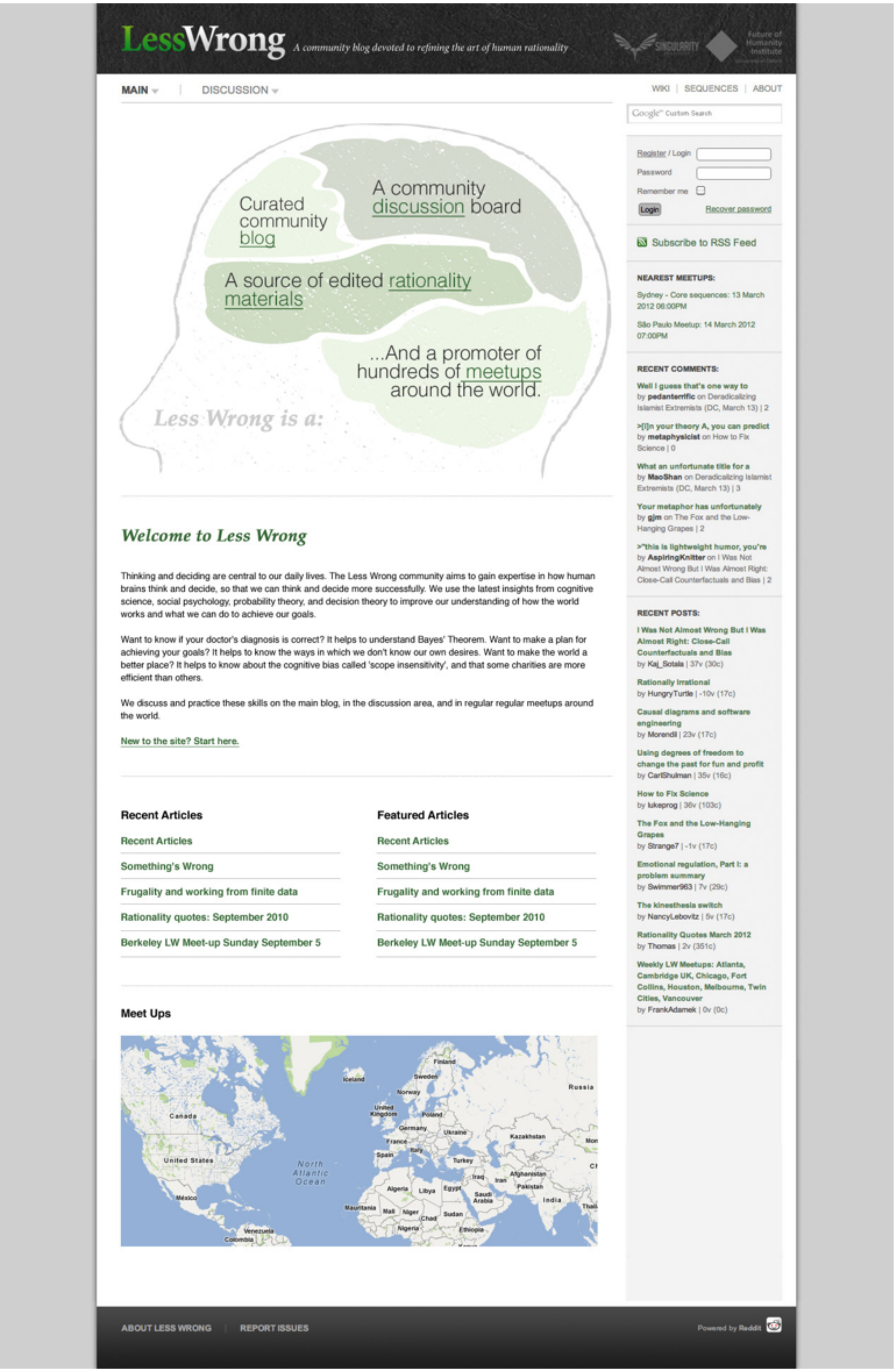
Then Critch broke up the class into pairs and asked each pair to carry out a similar operation on each other: Pick two plausible objects and make sure you're holding at least one of them, touch it to the other person while they face the other way, prior odds, additional fact, likelihood ratio, posterior odds.

This is the sort of in-person, hands-on, real-life, and *social* exercise that didn't occur to me, or Anna, or anyone else helping, while we were trying to design the Bayes's Theorem unit. Our brains just didn't go in that direction, though we recognized it as embarrassingly obvious in retrospect.

So... how would you design an exercise to teach Checking Consequentialism?

What if the front page…

What if the front page looked a little more like this?



LessWrong *A community blog devoted to refining the art of human rationality*

Curated community blog

A community discussion board

A source of edited rationality materials

...And a promoter of hundreds of meetups around the world.

Less Wrong is a:

Welcome to Less Wrong

Thinking and deciding are central to our daily lives. The Less Wrong community aims to gain expertise in how human brains think and decide, so that we can think and decide more successfully. We use the latest insights from cognitive science, social psychology, probability theory, and decision theory to improve our understanding of how the world works and what we can do to achieve our goals.

Want to know if your doctor's diagnosis is correct? It helps to understand Bayes' Theorem. Want to make a plan for achieving your goals? It helps to know the ways in which we don't know our own desires. Want to make the world a better place? It helps to know about the cognitive bias called 'scope insensitivity', and that some charities are more efficient than others.

We discuss and practice these skills on the main blog, in the discussion area, and in regular regular meetups around the world.

New to the site? Start here.

Recent Articles	Featured Articles
Recent Articles	Recent Articles
Something's Wrong	Something's Wrong
Frugality and working from finite data	Frugality and working from finite data
Rationality quotes: September 2010	Rationality quotes: September 2010
Berkeley LW Meet-up Sunday September 5	Berkeley LW Meet-up Sunday September 5

Meet Ups

Subscribe to RSS Feed

NEAREST MEETUPS:

Sydney - Core sequences: 13 March 2012 06:00PM

São Paulo Meetup: 14 March 2012 07:00PM

RECENT COMMENTS:

Well I guess that's one way to
by **pedanterrific** on Deradicalizing Islamist Extremists (DC, March 13) | 2

>[I]n your theory A, you can predict
by **metaphysicist** on How to Fix Science | 0

What an unfortunate title for a
by **MaoShan** on Deradicalizing Islamist Extremists (DC, March 13) | 3

Your metaphor has unfortunately
by **gjm** on The Fox and the Low-Hanging Grapes | 2

>"this is lightweight humor, you're
by **AspiringKnitter** on I Was Not Almost Wrong But I Was Almost Right: Close-Call Counterfactuals and Bias | 2

RECENT POSTS:

I Was Not Almost Wrong But I Was Almost Right: Close-Call Counterfactuals and Bias
by Kaj_Sotala | 37v (30c)

Rationally Irrational
by HungryTurtle | -10v (17c)

Causal diagrams and software engineering
by Morendil | 23v (17c)

Using degrees of freedom to change the past for fun and profit
by CarlShulman | 35v (16c)

How to Fix Science
by lukeprog | 36v (103c)

The Fox and the Low-Hanging Grapes
by Strange7 | -1v (17c)

Emotional regulation, Part I: a problem summary
by Swimmer963 | 7v (29c)

The kinesthesia switch
by NancyLebovitz | 5v (17c)

Rationality Quotes March 2012
by Thomas | 2v (351c)

Weekly LW Meetups: Atlanta, Cambridge UK, Chicago, Fort Collins, Houston, Melbourne, Twin Cities, Vancouver
by FrankAdamek | 0v (0c)

(Please assume that I'm trying to help. If you're polite and constructive (even if you hate this design) Omega will send you bundles of love in the post. If you're rude, I'll personally fund ninja assassins to hunt you down.)

Social status hacks from The Improv Wiki

I can't remember how I found this, just that I was amazed at how rational and near-mode it is on a topic where most of the information one usually encounters is hopelessly far.

LessWrong wiki link on the same topic: http://wiki.lesswrong.com/wiki/Status

The Improv Wiki

Status

Status is pecking order. The person who is lower in status defers to the person who is higher in status.

Status is party established by social position--e.g. boss and employee--but mainly by the way you interact. If you interact in a way that says you are not to be trifled with, the other person must adjust to you, then you are establishing high status. If you interact in a way that says you are willing to go along, you don't want responsibility, that's low status. A boss can play low status or high status. An employee can play low status or high status.

Status is established in every line and gesture, and changes continuously. Status is something that one character plays *to* another at a particular moment. If you convey that the other person must not cross you on what you're saying now, then you are playing high status to that person in that line. Your very next line might come out low status, as you suggest willingness to defer about something else.

If you analyze your most successful scenes, it's likely they involved several status changes between the players. Therefore, one path to great scenes is to intentionally change status. You can raise or lower your own status, or the status of the other player. The more subtly you can do this, the better the scene.

High-status behaviors

When walking, assuming that other people will get out of your path.

Making eye contact while speaking.

Not checking the other person's eyes for a reaction to what you said.

Having no visible reaction to what the other person said. (Imagine saying something to a typical Clint Eastwood character. You say something expecting a reaction, and you get--nothing.)

Speaking in complete sentences.

Interrupting before you know what you are going to say.

Spreading out your body to full comfort. Taking up a lot of space with your body.

Looking at the other person with your eyes somewhat down (head tilted back a bit to make this work), creating the feeling that you are a parent talking to a child.

Talking matter-of-factly about things that the other person finds displeasing or offensive.

Letting your body be vulnerable, exposing your neck and torso to the other person.

Moving comfortably and gracefully.

Keeping your hands away from your face.

Speaking authoritatively, with certainty.

Making decisions for a group; taking responsibility.

Giving or withholding permission.

Evaluating other people's work.

Speaking cryptically, not adjusting your speech to be easily understood by the other person (except that mumbling does not count). E.g. saying, "Chomper not right" with no explanation of what you mean or what you want the other person to do.

Being surrounded by an entourage, especially of people who are physically smaller than you.

A "high-status specialist" conveys in every word and gesture, "Don't come near me, I bite."

Low-status behaviors

When walking, moving out of other people's path.

Looking away from the other person's eyes.

Briefly checking the other person's eyes to see if they reacted positively to what you said.

Speaking in halting, incomplete sentences. Trailing off, editing your sentences as you got.

Sitting or standing uncomfortably in order to adjust to the other person and give them space. Pulling inward to give the other person more room. If you're tall, you might need to scrunch down a bit to indicate that you're not going to use your height against the other person.

Looking up toward the other person (head tilted forward a bit to make this work), creating the feeling that you are a child talking to a parent.

Dancing around your words (beating around the bush) when talking about something that will displease the other person.

Shouting as an attempt to intimidate the other person. This is low status because it suggests that you expect resistance.

Crouching your body as if to ward off a blow; protecting your face, neck, and torso.

Moving awkwardly or jerkily, with unnecessary movements.

Touching your face or head.

Avoiding making decisions for the group; avoiding responsibility.

Needing permission before you can act.

Adjusting the way you say something to help the other person understand; meeting the other person on their (cognitive) ground; explaining yourself. E.g. "Could you please adjust the chomper? That's the gadget on the kitchen counter immediately to the left of the toaster. If you just give it a slight rap on the top, that should adjust it."

A "low-status specialist" conveys in every word and gesture, "Please don't bite me, I'm not worth the trouble."

Raising another person's status

To raise another person's status is to establish them as high in the pecking order in your group (possibly just the two of you).

- Ask their permission to do something.
- Ask their opinion about something.
- Ask them for advice or help.
- Express gratitude for something they did.
- Apologize to them for something you did.
- Agree that they are right and you were wrong.
- Defer to their judgement without requiring proof.
- Address them with a fancy title or honorific (even "Mr." or "Sir" works very well).
- Downplay your own achievement or attribute in comparison to theirs. "Your wedding cake is so much whiter than mine."
- Do something incompetent in front of them and then apologize for it or act sheepish about it.
- Mention a failure or shortcoming of your own. "I was supposed to go to an audition today, but I was late. They said I was wrong for the part anyway."
- Compliment them in a way that suggests appreciation, not judgement. "Wow, what a beautiful cat you have!"
- Obey them unquestioningly.

- Back down in a conflict.
- Move out of their way, bow to them, lower yourself before them.
- Tip your hat to them.
- Lose to them at something competitive, like a game (or any comparison).
- Wait for them.
- Serve them; do manual labor for them.

Tip: Whenever you bring an audience member on stage, always raise their status, never lower it.

Lowering another person's status

To lower another person's status is to attack or discredit their right to be high in the pecking order. Another word for "lowering someone's status" is "humiliating them."

- Criticize something they did.
- Contradict them. Tell them they are wrong. Prove it with facts and logic.
- Correct them.
- Insult them.
- Give them unsolicited advice.
- Approve or disapprove of something they did or some attribute of theirs. "Your cat has both nose and ear points. That is acceptable." Anything that sets you up as the judge lowers their status, even "Nice work on the Milligan account, Joe."
- Shout at them.
- Tell them what to do.
- Ignore what they said and talk about something else, especially when they've said something that requires an answer. E.g. "Have you seen my socks?" "The train leaves in five minutes."
- One-up them. E.g. have a worse problem than the one they described, have a greater past achievement than theirs, have met a more famous celebrity, earn more money, do better than them at something they're good at, etc.
- Win: beat them at something competitive, like a game (or any comparison).
- Announce something good about yourself or something you did. "I went to an audition today, and I got the part!"
- Disregard their opinion. E.g. "You'd better not smoke while pumping gas, it's a fire hazard." Flick, light, puff, puff, pump, pump.
- Talk sarcastically to them.
- Make them wait for you.
- When they've fallen behind you, don't wait for them to catch up, just push on and get further out of sync.
- Disobey them.

- Violate their space.
- Beat them up. Beating them up verbally, not physically as in martial arts or how you learned UFC fighting in an gym, in front of other people, especially their wife, girlfriend, and/or children, is particularly status-lowering.
- In a conflict, make them back down.
- Taunt them. Tease them.

The basic status-lowering act

Laugh at them. (Not with them.)

The basic status-raising act

Be laughed at by them.

Second to that is laughing with them at someone else.

(Notice that those are primarily what comedians do.)

Note that behaviors that raise another person's status are not necessarily low-status behaviors, and behaviors that lower another person's status are not necessarily high-status behaviors. People at any status level raise and lower each other all the time. They can do so in ways that convey high or low status.

For example, shouting at someone lowers their status but is itself a low-status behavior.

Objects and environments also have high or low status, although this is seldom explored. So explore it. Make something cheap and inconsequential high status. (This fingernail clipping came from Graceland!) Or bring down the status of a high status item. (Casually toss a 2 carat diamond ring on your jewelry pile.)

Source: http://greenlightwiki.com/improv/Status
Retrieved 20 March 2012

Causal diagrams and software engineering

> Fake explanations don't feel fake. That's what makes them dangerous. -- EY

Let's look at "A Handbook of Software and Systems Engineering", which purports to examine the insights from software engineering that are *solidly* grounded in empirical evidence. Published by the prestigious Fraunhofer Institut, this book's subtitle is in fact "Empirical Observations, Laws and Theories".

Now "law" is a strong word to use - the highest level to which an explanation can aspire to reach, as it were. Sometimes it's used in a jokey manner, as in "Hofstadter's Law" (which certainly *seems* often to apply to software projects). But this definitely isn't a jokey kind of book, that much we get from the appeal to "empirical observations" and the "handbook" denomination.

Here is the very first "law" listed in the Handbook:

> Requirement deficiencies are the prime source of project failures.

Previously, we observed that in the field of software engineering, a last name followed by a year, surrounded by parentheses, seems to be a magic formula for suspending critical judgment in readers.

Another such formula, it seems, is the invocation of statistical results. Brandish the word "percentage", assert that you have surveyed a largish population, and whatever it is you claim, *some* people will start believing. Do it often enough and some will start repeating your claim - without bothering to check it - starting a potentially viral cycle.

As a case in point, one of the most often cited pieces of "evidence" in support of the above "law" is the well-known Chaos Report, according to which the first cause of project failure is "Incomplete Requirements". (The Chaos Report isn't cited as evidence by the Handbook, but it's representative enough to serve in the following discussion. A Google Search readily attests to the wide spread of the *verbatim* claim in the Chaos Report; various derivatives of the claim are harder to track, but easily verified to be quite pervasive.)

Some elementary reasoning about causal inference is enough to show that the same evidence supporting the above "law" can equally well be suggested as evidence supporting this alternative conclusion:

> Project failures are the primary source of requirements deficiencies.

"Wait", you may be thinking. "Requirements are written at the start of a project, and the outcome (success or failure) happens at the end. The latter cannot be the cause of the former!"

Your thinking is correct! As the descendant of a long line of forebears who, by virtue of observing causes and effects, avoided various dangers such as getting eaten by predators, you have internalized a number of constraints on causal inference. Without necessarily having an explicit representation of these constraints, you know that at a minimum, showing a cause-effect relationship requires the following:

- a relationship between the variable labeled "cause" and the variable labeled "effect"; it need not be deterministic (as in "Y always happens after X") but can also be probabilistic ("association" or "correlation")
- the cause must have happened before the effect
- other causes which could also explain the effect are ruled out by reasoning or observation

Yet, notoriously, we often fall prey to the failure mode of only requiring the first of these conditions to be met:

> Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'.

One of the more recent conceptual tools for avoiding this trap is to base one's reasoning on formal representations of cause-effect inferences, which can then serve to suggest the quantitative relationships that will confirm (or invalidate) a causal claim. Formalizing helps us bring to bear all that we know about the structure of reliable causal inferences.

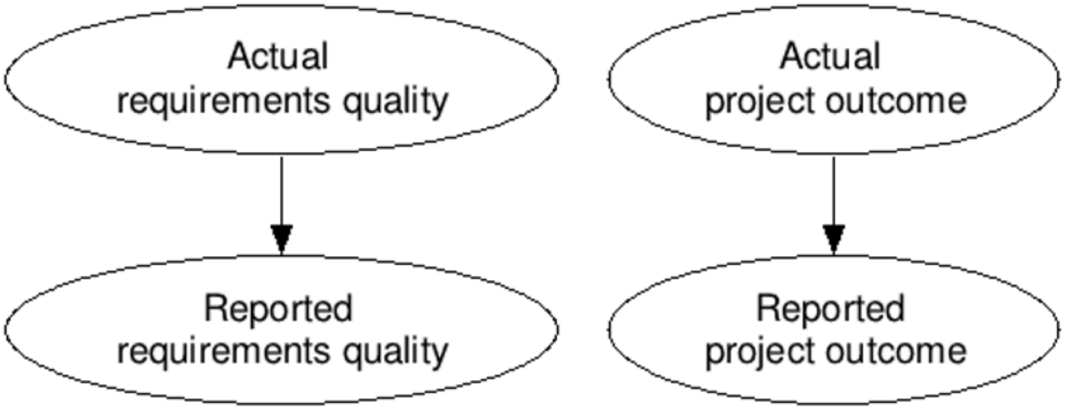
The "ruling out alternate explanations" bit turns out to be kind of a big deal. It is, in fact, a large part of the difference between "research" and "anecdote". The reason you can't stick the 'science' label on everyday observations isn't generally because these are imprecise and merely qualitative, and observing percentages or averages is sufficient to somehow obtain *science*.

There are no mathematical operations which magically transform observations into valid inferences. Rather, to do "science" consists in good part of eliminating the various ways you could be fooling yourself. The hard part isn't collecting the data; the hard part is *designing* the data collection, so that the data actually tells you something useful.

Here is an elementary practical application, in the context of the above "law"; let's look at the study design. What is the Chaos Report's methodology for establishing the widely circulated list of "top causes of software project failure"?

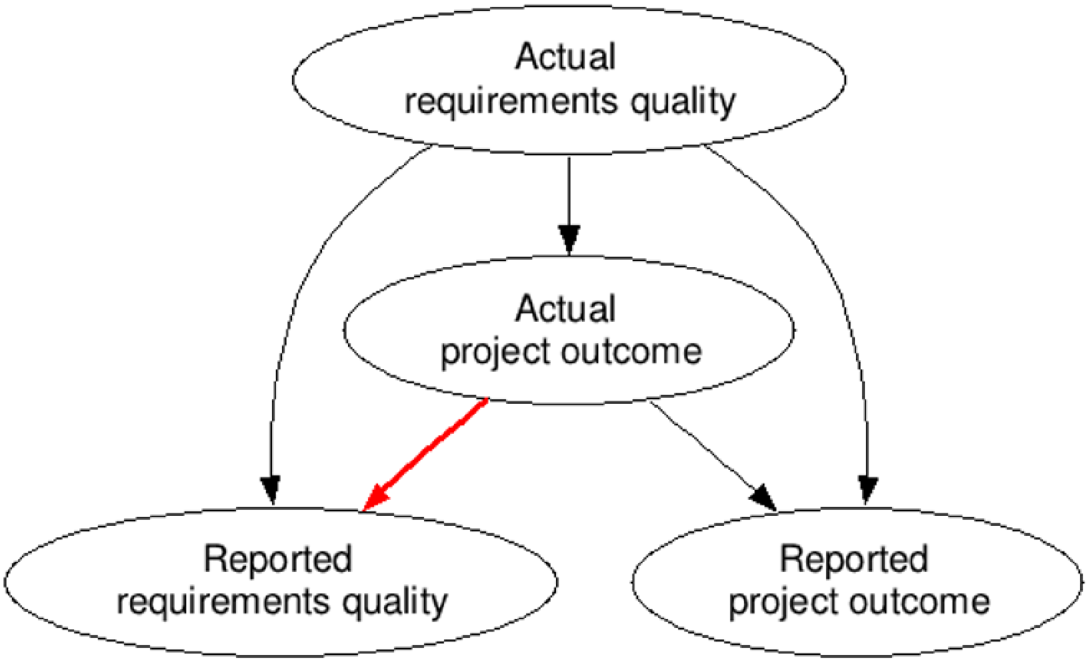
> The Standish Group surveyed IT executive managers for their opinions about why projects succeed. (source)
>
> The respondents to the Standish Group survey were IT executive managers. The sample included large, medium, and small companies across major industry segments: banking, securities, manufacturing, retail, wholesale, heath care, insurance, services, local, state, and federal organizations. The total sample size was 365 respondents representing 8,380 applications.

The key terms here are "survey" and "opinion". IT executives are being interviewed, after the relevant projects have been conducted and assessed, on what they think best explains the outcomes. We can formalize this with a causal diagram. We need to show four variables, and there are some obvious causal relationships:



Note that in a survey situation, the values of the "actual" variables are not measured directly; they are only "measured" indirectly via their effects on the "reported" variables.

As surmised above, we may rule out any effect of the reported results on the actual results, since the survey takes place after the projects. However, we may *not* rule out an effect of the actual results on the reported results, in either direction. This is reflected in our diagram as follows:



An argument for the arrow in red could be formulated this way: "An IT executive being interviewed about the reason for a project failure is less likely to implicate his own competence, and more likely to implicate the competence of some part of the organization outside of their scope of responsibility, for instance by blaming his non-IT interlocutors for poor requirements definition or insufficient involvement." This isn't just possible, it's also plausible (based on what we know of human nature and corporate politics).

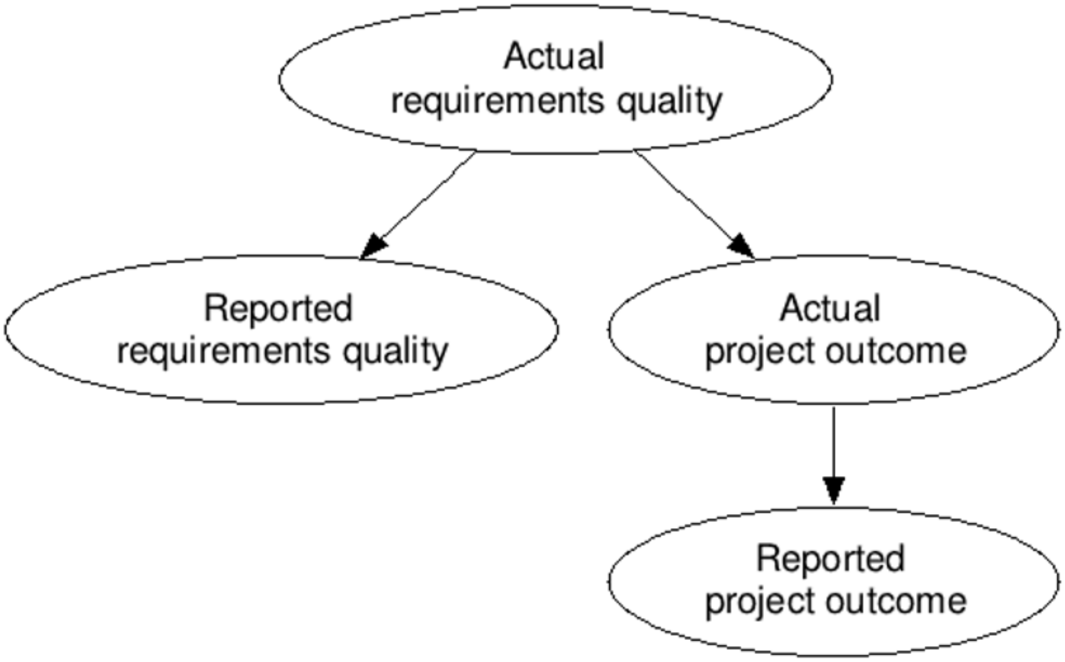
Hence my claim above: we can equally well argue from the evidence that "(actual) project outcomes are the primary source of (reported) requirements deficiencies".

This is another piece of critical kit that a rationalist plying their trade in the software development business (and indeed, a rationalist anywhere) cannot afford to be without: correctly generating and labeling (if only mentally) the nodes and edges in an implied causal diagram, whenever reading about the results of an experimental or observational study. (The diagrams for this post were created using the nifty GraphViz Workspace.)

We might even see it as a candidate 5-second skill, which unlocks the really powerful habit: asking the question "which causal pathways between cause and effect, that yield an alternative explanation to the hypothesis under consideration, could be ruled out by an alternative experimental design?"

Sometimes it's hard to come up with a suitable experimental design, and in such cases there are sophisticated mathematical techniques emerging that let you *still* extract good causal inferences from the imperfect data. This... isn't one of those cases.

For instance, a differently designed survey would interview IT executives at the *start* of every project, and ask them the question: "do you think your user-supplied requirements are complete enough that this will not adversely impact your project?" As before, you would debrief the same executives at the end of the project. This study design yields the following causal diagram:



Observe that we have now ruled out the argument from CYA. This is a simple enough fix, yet the industry for the most part (and despite sporadic outbreaks of common sense) persists in conducting and uncritically quoting surveys that do not block enough causal pathways to firmly establish the conclusions they report, conclusions that have been floating around, for the most part uncontested, for decades now.

Doing "Nothing"

It might be a useful habit to remember, whenever you're making a choice about some situation, that "doing nothing" is never actually an available option. Even if you avoid doing the task you're considering, you're still making some kind of choice about how you spend your time, and you're still doing something relative to that task. For example, if the task is "paint the barn" the alternative is "leave the bare barn exposed to the elements", not "store the barn in some impermeable stasis field and return to paint it later". Being able to clearly articulate what that "nothing" slot entails, its consequences and rewards, might be a helpful way to motivate yourself to make better choices.

I am working on internalising this, because if I don't think about it, a part of me tends to just think that I'm doing the equivalent of sticking the task in an atemporal stasis field instead of leaving it unattended. If I don't exercise, I don't stay "the same amount fit". I get *weaker* (or, as **aelephant** points out, I could be getting stronger, during a recovery period - in which case "doing nothing" (as far as exercise) is the *better* option, after evaluation) . If I don't study, I don't stay "the same amount knowledgeable". *I forget*. Sure, there are things which remain effectively "in stasis" - Olympus Mons will probably stay about the same whether I climb it in ten years (somehow) or a hundred years - but I won't be the same by then. Or things that are so transient and commonplace that they might as well be in stasis - If I'm thinking of going somewhere, I might think, "I might miss catching this taxi cab, but I miss cabs all the time, there are always more cabs, and I can catch another one". But subjectively static opportunities are rare.

6 Tips for Productive Arguments

We've all had arguments that seemed like a complete waste of time in retrospect. But at the same time, arguments (between scientists, policy analysts, and others) play a critical part in moving society forward. You can imagine how lousy things would be if no one ever engaged those who disagreed with them.

This is a list of tips for having "productive" arguments. For the purposes of this list, "productive" means improving the accuracy of at least one person's views on some important topic. By this definition, arguments where no one changes their mind are unproductive. So are arguments about unimportant topics like which Pink Floyd album is the best.

Why do we want productive arguments? Same reason we want Wikipedia: so people are more knowledgeable. And just like the case of Wikipedia, there is a strong selfish imperative here: arguing *can* make you more knowledgeable, if you're willing to change your mind when another arguer has better points.

Arguments can also be *negatively* productive if everyone moves *further* from the truth on net. This could happen if, for example, the truth was somewhere in between two arguers, but they both left the argument even more sure of themselves.

These tips are derived from my personal experience arguing.

Keep it Friendly

Probably the biggest barrier to productive arguments is the desire of arguers to save face and avoid publicly admitting they were wrong. Obviously, it's hard for anyone's views to get more accurate if no one's views ever change.

This problem is exacerbated when arguers disparage one another. If you rebuke a fellow arguer, you're setting yourself up as their enemy. Admitting they were wrong would then mean giving in to an enemy. And no one likes to do that.
You may also find it difficult to carefully reconsider your *own* views after having ridiculed or berated someone who disagrees. I know I have in the past.
Both of these tendencies hurt argument productivity. To make arguments productive:

- Keep things warm and collegial. Just because your ideas are in violent disagreement doesn't mean you have to disagree violently as people. Stay classy.
- To the greatest extent possible, uphold the social norm that no one will lose face for publicly changing their mind.
- If you're on a community-moderated forum like Less Wrong, don't downvote something unless you think the person who wrote it is being a bad forum citizen (ex: spam or unprovoked insults). Upvotes already provide plenty of information about how comments and submissions should be sorted. (It's probably safe to assume that a new Less Wrong user who sees their first comment modded below zero will decide we're all jerks and never come back. And if new users aren't coming back, we'll have a hard time raising the sanity waterline much.)

- Err on the side of understating your disagreement, e.g. "I'm not persuaded that..." or "I agree that x is true; I'm not as sure that..." or "It seems to me..."
- If you notice some hypocrisy, bias, or general deficiency on the part of another arguer, think extremely carefully before bringing it up while the argument is still in progress.

In a good argument, all parties will be curious about what's really going on. But curiosity and animosity are mutually incompatible emotions. Don't impede the collective search for truth through rudeness or hostility.

Inquire about Implausible-Sounding Assertions Before Expressing an Opinion

It's easy to respond to a statement you think is obviously wrong with with an immediate denial or attack. But this is also a good way to keep yourself from learning anything.

If someone suggests something you find implausible, start asking friendly questions to get them to clarify and justify their statement. If their reasoning seems genuinely bad, you can refute it then.

As a bonus, doing nothing but ask questions can be a good way to save face if the implausible assertion-maker turns out to be right.

Be careful about rejecting highly implausible ideas out of hand. Ideally, you want your rationality to be a level where even if you started out with a crazy belief like Scientology, you'd still be able to get rid of it. But for a Scientologist to berid themselves of Scientology, they have to consider ideas that initially seen extremely unlikely.

It's been argued that many mainstream skeptics aren't really that good at critically evaluating ideas, just dismissing ones that seem implausible.

Isolate Specific Points of Disagreement

Stick to one topic at a time, until someone changes their mind or the topic is declared not worth pursuing. If your discussion constantly jumps from one point of disagreement to another, reaching consensus on anything will be difficult.

You can use hypothetical-oriented thinking like conditional probabilities and the least convenient possible world to figure out exactly what it is you disagree on with regard to a given topic. Once you've creatively helped yourself or another arguer clarify beliefs, sharing intuitions on specific "irreducible" assertions or anticipated outcomes that aren't easily decomposed can improve both of your probability estimates.

Don't Straw Man Fellow Arguers, Steel Man Them Instead

You might think that a productive argument is one where the smartest person wins, but that's not always the case. Smart people can be wrong too. And a smart person successfully convincing less intelligent folks of their delusion counts as a negatively productive argument (see definition above).

Play for all sides, in case you're the smartest person in the argument.

Rewrite fellow arguers' arguments so they're even stronger, and think of new ones. Arguments for new positions, even—they don't have *anyone* playing for them. And if you end up convincing yourself of something you didn't previously believe, so much the better.

If You See an Opportunity To Improve the Accuracy of Your Knowledge, Take It!

This is often called losing an argument, but you're actually the winner: you and your arguing partner both invested time to argue, but you were the only one who received significantly improved knowledge.

I'm not a Christian, but I definitely want to know if Christianity is true so I can stop taking the Lord's name in vain and hopefully get to heaven. (Please don't contact me about Christianity though, I've already thought about it a lot and judged it too improbable to be worth spending additional time thinking about.) Point is, it's hard to see how having more accurate knowledge could *hurt*.

If you're worried about losing face or seeing your coalition (research group, political party, etc.) diminish in importance from you admitting that you were wrong, here are some ideas:

- Say "I'll think about it". Most people will quiet down at this point without any gloating.
- Just keep arguing, making a mental note that your mind has changed.
- Redirect the conversation, pretend to lose interest, pretend you have no time to continue arguing, etc.

If necessary, you can make up a story about how something else changed your mind later.

Some of these techniques may seem dodgy, and honestly I think you'll usually do better by explaining what actually changed your mind. But they're a small price to pay for more accurate knowledge. Better to tell unimportant false statements to others than important false statements to yourself.

Have Low "Belief Inertia"

It's actually pretty rare that the evidence that you're wrong comes suddenly—usually you can see things turning against you. As an advanced move, cultivate the ability to update your degree of certainty in real time to new arguments, and tell fellow arguers if you find an argument of theirs persuasive. This can actually be a good way to make friends. It also encourages other arguers to share additional arguments with you, which could be valuable data.

One psychologist I agree with suggested that people ask

- "Does the evidence **allow** me to believe?" when evaluating what they already believe, but
- "Does the evidence **compel** me to believe?" when evaluating a claim incompatible with their current beliefs.

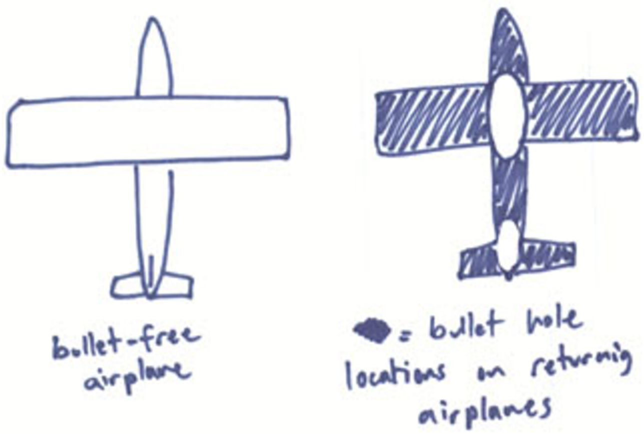
If folks don't have to drag you around like this for you to change your mind, you don't actually lose much face. It's only long-overdue capitulations that result in significant face loss. And the longer you put your capitulation off, the worse things get. Quickly updating in response to new evidence seems to *preserve* face in my experience.

If your belief inertia is low and you steel-man everything, you'll reach the super chill state of not having a "side" in any given argument. You'll play for all sides and you won't care who wins. You'll have achieved equanimity, content with the world as it actually is, not how you wish it was.

Examine your assumptions

There's a story you've probably heard:

> During World War II, the British RAF's Bomber Command wanted a survey done on the effectiveness of their aircraft armouring. This was carried out by inspected all bombers returning from bombing raids over Germany over a particular period. All damage inflicted by German air defences was noted and the recommendation was given that armour be added in the most heavily damaged areas.



> However a new group, run by Patrick Blackett, the Operational Research Section, analysed the survey report, and came to a different conclusion. Blackett suggested that, instead, the armour be placed in the areas which were completely untouched by damage in the bombers which returned. He reasoned that the survey was biased, since it only included aircraft that returned to Britain. The untouched areas of returning aircraft were probably vital areas, which, if hit, would result in the loss of the aircraft.

It is a useful fable, but in the context presented it seemed unlikely, given the attitudes of Bomber Harris. So I went looking for further information, and found the story of BC-ORS written by Freeman Dyson:

"A Failure of Intelligence" (Part 1) (Part 2)

which is a great read, but fails to mention any such incident.

What I did find, however, on further searching, was the work of Abraham Wald. Wald was a Jewish mathematician from Romania who in 1943 published a series of 8 memoranda via the Statistical Research Group at Columbia University while working for the National Defense Research Committee in America. These were republished collectively in 1980 as "'A Method of Estimating Plane Vulnerability Based on Damage of Survivors." by the Center for Naval Analyses, and are still in use today.

In 1984 Mangel and Samaniego published a fairly accessible summary of Wald's work in the Journal of the American Statistical Association (Vol 79, Issue 286, June)

"Abraham Wald's Work on Aircraft Survivability"

So it seems that Wald is the one who should get the credit for being the first to try to compensate for the evidential problem. Tragically he himself died in an airplane crash, just a few years later (in 1950, aged 48).

The 'bible' on this topic, Robert Ball's "The Fundamentals of Aircraft Combat Survivability Analysis and Design" confirms the problem is a real one, and mentioned the F-4 as an example. When they looked at the F-4s which survived combat, there were no holes in the narrowest part of the tail, just forward of the horizontal stabilizers. They figured out that all of the hydraulic lines for the elevators and rudder were tightly clustered in there, so that a single hit could damage all of them at once, leaving the plane uncontrollable. The solution in that case was, rather than increasing the armour, to spread the redundant lines out to reduce the chances of losing all of them to a single hit.