Lesswrong Core Sequences

Elizier Yudkowsky

Contents

How To Actually Change Your Mind

Politics is the Mind-Killer

Death Spirals and the Cult Attractor

Seeing with Fresh Eyes

Against Rationalization

Against Doublethink

Overly Convenient Excuses

Letting Go

Part I
Map and Territory

A collection of posts dealing with the fundamentals of rationality: the difference between the map and the territory, Bayes's Theorem and the nature of evidence, why anyone should care about truth, and minds as reflective cognitive engines.

1. What Do We Mean By "Rationality"?[1]

We mean:

1. **Epistemic rationality**: believing, and updating on evidence, so as to systematically improve the correspondence between your map and the territory[2]. The art of obtaining beliefs that correspond to reality as closely as possible. This correspondence is commonly termed "truth" or "accuracy", and we're happy to call it that.
2. **Instrumental rationality**: achieving your values. *Not* necessarily "your values" in the sense of being *selfish* values or *unshared* values: "your values" means *anything you care about*. The art of choosing actions that steer the future toward outcomes ranked higher in your preferences. On LW we sometimes refer to this as "winning".

If that seems like a perfectly good definition, you can stop reading here; otherwise continue.

Sometimes experimental psychologists uncover human reasoning that seems very strange - for example[3], someone rates the probability "Bill plays jazz" as *less* than the probability "Bill is an accountant who plays jazz". This seems like an odd judgment, since any particular jazz-playing accountant is obviously a jazz player. But to what higher vantage point do we appeal in saying that the judgment is *wrong?*

Experimental psychologists use two gold standards: *probability theory,* and *decision theory*. Since it is a universal law of probability theory that P(A) ≥ P(A & B), the judgment P("Bill plays jazz") < P("Bill plays jazz" & "Bill is accountant") is labeled incorrect.

---

1. *http://lesswrong.com/lw/31/what_do_we_mean_by_rationality/*
2. *http://yudkowsky.net/rational/the-simple-truth*
3. *http://lesswrong.com/lw/ji/conjunction_fallacy/*

To keep it technical, you would say that this probability judgment is *non-Bayesian*. Beliefs that conform to a coherent probability distribution, and decisions that maximize the probabilistic expectation of a coherent utility function, are called "Bayesian".

This does not quite exhaust the problem of what is meant in practice by "rationality", for two major reasons:

First, the Bayesian formalisms in their full form are computationally intractable on most real-world problems. No one can *actually* calculate and obey the math, any more than you can predict the stock market by calculating the movements of quarks.

This is why we have a whole site called "Less Wrong", rather than simply stating the formal axioms and being done. There's a whole further art to finding the truth and accomplishing value *from inside a human mind:* we have to learn our own flaws, overcome our biases, prevent ourselves from self-deceiving, get ourselves into good emotional shape to confront the truth and do what needs doing, etcetera etcetera and so on.

Second, sometimes the meaning of the math itself is called into question. The exact rules of probability theory are called into question by e.g. anthropic problems[4] in which the number of observers is uncertain. The exact rules of decision theory are called into question by e.g. Newcomblike problems[5] in which other agents may predict your decision before it happens.

In cases like these, it is futile to try to settle the problem by coming up with some new definition of the word "rational", and saying, "Therefore my preferred answer, *by definition,* is what is meant by the word 'rational'." This simply begs the question of why anyone should pay attention to your definition. We aren't interested in probability theory because it is the holy word handed down from Laplace. We're interested in Bayesian-style belief-updating (with Occam priors) because we

4. *http://www.anthropic-principle.com/primer.html*
5. *http://lesswrong.com/lw/nc/newcombs_problem_and_regret_of_rationality/*

expect that this style of thinking gets us systematically closer to, you know, *accuracy,* the map that reflects the territory. (More on the futility of arguing "by definition" here[6] and here[7].)

And then there are questions of "How to think" that seem not quite answered by either probability theory or decision theory - like the question of how to feel about the truth once we have it[8]. Here again, trying to define "rationality" a particular way doesn't support an answer, merely presume it.

From the Twelve Virtues of Rationality[9]*:*

> How can you improve your conception of rationality? Not by saying to yourself, "It is my duty to be rational." By this you only enshrine your mistaken conception. Perhaps your conception of rationality is that it is rational to believe the words of the Great Teacher, and the Great Teacher says, "The sky is green," and you look up at the sky and see blue. If you think: "It may look like the sky is blue, but rationality is to believe the words of the Great Teacher," you lose a chance to discover your mistake.
>
> Do not ask whether it is "the Way" to do this or that. Ask whether the sky is blue or green. If you speak overmuch of the Way you will not attain it.
>
> You may try to name the highest principle with names such as "the map that reflects the territory" or "experience of success and failure" or "Bayesian decision theory". But perhaps you describe incorrectly the nameless virtue. How

6. *http://lesswrong.com/lw/nf/the_parable_of_hemlock/*
7. *http://lesswrong.com/lw/nz/arguing_by_definition/*
8. Page 463, 'Feeling Rational'.
9. *http://yudkowsky.net/virtues/*

> will you discover your mistake? Not by comparing your description to itself, but by comparing it to that which you did not name.

We are not here to argue the meaning of a word[10], not even if that word is "rationality". The point of attaching sequences of letters to particular concepts is to let two people *communicate*[11] - to help transport thoughts from one mind to another. You cannot change reality, or prove the thought, by manipulating which meanings go with which words.

So if you understand what concept we are *generally getting at* with this word "rationality", and with the sub-terms "epistemic rationality" and "instrumental rationality", we *have communicated:* we have accomplished everything there is to accomplish by talking about how to define "rationality". What's left to discuss is not *what meaning* to attach to the syllables "ra-tio-na-li-ty"; what's left to discuss is *what is a good way to think*.

With that said, you should be aware that many of us will regard as *controversial* - at the very least - any construal of "rationality" that makes it *non-normative:*

For example, if you say, "The rational belief is X, but the true belief is Y" then you are probably using the word "rational" in a way that means something other than what most of us have in mind. (E.g. some of us expect "rationality" to be *consistent under reflection* - "rationally" looking at the evidence, and "rationally" considering how your mind processes the evidence, shouldn't lead to two different conclusions.) Similarly, if you find yourself saying "The rational thing to do is X, but the right thing to do is Y" then you are almost certainly using one of the words "rational" or "right" in a way that a huge chunk of readers won't agree with.

10. *http://lesswrong.com/lw/np/disputing_definitions/*
11. *http://lesswrong.com/lw/nr/the_argument_from_common_usage/*

In this case - or in any other case where controversy threatens - you should substitute more specific language[12]: "The self-benefiting thing to do is to run away, but I hope I would at least try to drag the girl off the railroad tracks" or "Causal decision theory as usually formulated says you should two-box on Newcomb's Problem[13], but I'd rather have a million dollars."

"X is rational!" is usually just a more strident way of saying "I think X is true" or "I think X is good". So why have an additional word for "rational" as well as "true" and "good"? Because we want to talk about *systematic methods* for obtaining truth and winning.

The word "rational" has potential pitfalls, but there are plenty of *non*-borderline cases where "rational" works fine to *communicate* what one is getting at, likewise "irrational". In these cases we're not afraid to use it.

Yet one should also be careful not to *overuse* that word. One receives no points merely for pronouncing it loudly. If you speak overmuch of the Way you will not attain it.

12. *http://lesswrong.com/lw/nu/taboo_your_words/*
13. *http://lesswrong.com/lw/nc/newcombs_problem_and_regret_of_rationality/*

2. Why truth? And...[1]

Some of the comments in this blog have touched on the question of why we ought to seek truth. (Thankfully not many have questioned what truth is[2].) Our shaping motivation for configuring our thoughts to rationality, which determines whether a given configuration is "good" or "bad", comes from whyever we wanted to find truth in the first place.

It is written: "The first virtue is curiosity." Curiosity is one reason to seek truth, and it may not be the only one, but it has a special and admirable purity. If your motive is curiosity, you will assign priority to questions according to how the questions, themselves, tickle your personal aesthetic sense. A trickier challenge, with a greater probability of failure, may be worth more effort than a simpler one, just because it is more fun.

Some people, I suspect, may object that curiosity is an emotion and is therefore "not rational". I label an emotion as "not rational" if it rests on mistaken beliefs, or rather, on irrational epistemic conduct: "If the iron approaches your face, and you believe it is hot, and it is cool, the Way opposes your fear. If the iron approaches your face, and you believe it is cool, and it is hot, the Way opposes your calm." Conversely, then, an emotion which is evoked by correct beliefs or epistemically rational thinking is a "rational emotion"; and this has the advantage of letting us regard calm as an emotional state, rather than a privileged default. When people think of "emotion" and "rationality" as opposed, I suspect that they are really thinking of System 1 and System 2—fast perceptual judgments versus slow deliberative judgments. Deliberative judgments aren't always true, and perceptual judgments aren't always false; so it is very important to distinguish that dichotomy from "rationality". Both systems can serve the goal of truth, or defeat it, according to how they are used.

---

1. *http://lesswrong.com/lw/go/why_truth_and/*
2. *http://sl4.org/wiki/TheSimpleTruth*

Besides sheer emotional curiosity, what other motives are there for desiring truth? Well, you might want to accomplish some specific real-world goal, like building an airplane, and therefore you need to know some specific truth about aerodynamics. Or more mundanely, you want chocolate milk, and therefore you want to know whether the local grocery has chocolate milk, so you can choose whether to walk there or somewhere else. If this is the reason you want truth, then the priority you assign to your questions will reflect the expected utility of their information—how much the possible answers influence your choices, how much your choices matter, and how much you expect to find an answer that changes your choice from its default.

To seek truth merely for its instrumental value may seem impure—should we not desire the truth for its own sake?—but such investigations are extremely important because they create an outside criterion of verification: if your airplane drops out of the sky, or if you get to the store and find no chocolate milk, it's a hint that you did something wrong. You get back feedback on which modes of thinking work, and which don't. Pure curiosity is a wonderful thing, but it may not linger too long on verifying its answers, once the attractive mystery is gone. Curiosity, as a human emotion, has been around since long before the ancient Greeks. But what set humanity firmly on the path of Science was noticing that certain modes of thinking uncovered beliefs that let us *manipulate the world*. As far as sheer curiosity goes, spinning campfire tales of gods and heroes satisfied that desire just as well, and no one realized that anything was wrong with that.

Are there motives for seeking truth besides curiosity and pragmatism? The third reason that I can think of is morality: You believe that to seek the truth is noble and important and worthwhile. Though such an ideal also attaches an intrinsic value to truth, it's a very different state of mind from curiosity. Being curious about what's behind the curtain doesn't feel the same as believing that you have a moral duty to look there.

In the latter state of mind, you are a lot more likely to believe that someone *else* should look behind the curtain, too, or castigate them if they deliberately close their eyes. For this reason, I would also label as "morality" the belief that truthseeking is pragmatically important *to society*, and therefore is incumbent as a duty upon all. Your priorities, under this motivation, will be determined by your ideals about which truths are most important (not most useful or most intriguing); or your moral ideals about when, under what circumstances, the duty to seek truth is at its strongest.

I tend to be suspicious of morality as a motivation for rationality, *not* because I reject the moral ideal, but because it invites certain kinds of trouble. It is too easy to acquire, as learned moral duties, modes of thinking that are dreadful missteps in the dance. Consider Mr. Spock of *Star Trek*, a naive archetype of rationality. Spock's emotional state is always set to "calm", even when wildly inappropriate. He often gives many significant digits for probabilities that are grossly uncalibrated. (E.g: "Captain, if you steer the Enterprise directly into that black hole, our probability of surviving is only 2.234%" Yet nine times out of ten the Enterprise is not destroyed. What kind of tragic fool gives four significant digits for a figure that is off by two orders of magnitude?) Yet this popular image is how many people conceive of the duty to be "rational"—small wonder that they do not embrace it wholeheartedly. To make rationality into a moral duty is to give it all the dreadful degrees of freedom of an arbitrary tribal custom. People arrive at the wrong answer, and then indignantly protest that they acted with propriety, rather than learning from their mistake.

And yet if we're going to *improve* our skills of rationality, go beyond the standards of performance set by hunter-gatherers, we'll need deliberate beliefs about how to think with propriety. When we write new mental programs for ourselves, they start out in System 2, the deliberate system, and are only slowly—if ever—trained into the neural circuitry that underlies System 1. So if there are certain kinds of thinking that we find we want to

avoid—like, say, biases—it will end up represented, within System 2, as an injunction not to think that way; a professed duty of avoidance.

If we want the truth, we can most effectively obtain it by thinking in certain ways, rather than others; and these are the techniques of rationality. Some of the techniques of rationality involve overcoming a certain class of obstacles, the biases...

(Continued in next post: "What's a bias, again?")

3. What is Evidence?[1]

> "The sentence 'snow is white' is *true* if and only if snow is white."
>
> —Alfred Tarski
>
> "To say of what is, that it is, or of what is not, that it is not, is *true*."
>
> —Aristotle, *Metaphysics IV*

If these two quotes don't seem like a sufficient definition of "truth", read this[2]. Today I'm going to talk about "evidence". (I also intend to discuss beliefs-of-fact, not emotions or morality, as distinguished here[3].)

Walking along the street, your shoelaces come untied. Shortly thereafter, for some odd reason, you start *believing* your shoelaces are untied. Light leaves the Sun and strikes your shoelaces and bounces off; some photons enter the pupils of your eyes and strike your retina; the energy of the photons triggers neural impulses; the neural impulses are transmitted to the visual-processing areas of the brain; and there the optical information is processed and reconstructed into a 3D model that is recognized as an untied shoelace. There is a sequence of events, a chain of cause and effect, within the world and your brain, by which you end up believing what you believe. The final outcome of the process is a state of *mind* which mirrors the state of your actual *shoelaces*.

What is *evidence?* It is an event entangled, by links of cause and effect, with whatever you want to know about. If the target of your inquiry is your shoelaces, for example, then the light entering your pupils is evidence entangled with your shoelaces. This should not be confused with the technical sense of "entanglement" used in physics—here I'm just talking about "entan-

1. *http://lesswrong.com/lw/jl/what_is_evidence/*
2. *http://sl4.org/wiki/TheSimpleTruth*
3. Page 463, 'Feeling Rational'.

glement" in the sense of two things that end up in correlated states because of the links of cause and effect between them.

Not every influence creates the kind of "entanglement" required for evidence. It's no help to have a machine that beeps when you enter winning lottery numbers, if the machine *also* beeps when you enter *losing* lottery numbers. The light reflected from your shoes would not be useful evidence about your shoelaces, if the photons ended up in the same physical state whether your shoelaces were tied or untied.

To say it abstractly: For an event to be *evidence about* a target of inquiry, it has to happen *differently* in a way that's entangled with the *different* possible states of the target. (To say it technically: There has to be Shannon mutual information between the evidential event and the target of inquiry, relative to your current state of uncertainty about both of them.)

Entanglement can be contagious *when processed correctly,* which is why you need eyes and a brain. If photons reflect off your shoelaces and hit a rock, the rock won't change much. The rock won't reflect the shoelaces in any helpful way; it won't be detectably different depending on whether your shoelaces were tied or untied. This is why rocks are not useful witnesses in court. A photographic film will contract shoelace-entanglement from the incoming photons, so that the photo can itself act as evidence. If your eyes and brain work correctly, *you* will become tangled up with your own shoelaces.

This is why rationalists put such a heavy premium on the paradoxical-seeming claim that a belief is only really worthwhile if you could, in principle, be persuaded to believe otherwise. If your retina ended up in the same state regardless of what light entered it, you would be blind. Some belief systems, in a rather obvious trick to reinforce themselves, say that certain beliefs are only really worthwhile if you believe them *unconditionally*— no matter what you see, no matter what you think. Your brain is supposed to end up in the same state regardless. Hence the phrase, "blind faith". If what you believe

doesn't depend on what you see, you've been blinded as effectively as by poking out your eyeballs.

If your eyes and brain work correctly, your beliefs will end up entangled with the facts. *Rational thought produces beliefs which are themselves evidence.*

If your tongue speaks truly, your rational beliefs, which are themselves evidence, can act as evidence for someone else. Entanglement can be transmitted through chains of cause and effect—and if you speak, and another hears, that too is cause and effect. When you say "My shoelaces are untied" over a cellphone, you're sharing your entanglement with your shoelaces with a friend.

Therefore rational beliefs are contagious, among honest folk who believe each other to be honest. And it's why a claim that your beliefs are *not* contagious—that you believe for private reasons which are not transmissible—is so suspicious. If your beliefs are entangled with reality, they *should* be contagious among honest folk.

If your model of reality suggests that the outputs of your thought processes should *not* be contagious to others, then your model says that your beliefs are not themselves evidence, meaning they are not entangled with reality. You should apply a reflective correction, and stop believing.

Indeed, if you *feel,* on a *gut* level, what this all *means,* you will *automatically* stop believing. Because "my belief is not entangled with reality" *means* "my belief is not accurate". As soon as you stop believing "'snow is white' is true", you should (automatically!) stop believing "snow is white", or something is very wrong.

So go ahead and explain why the kind of thought processes you use systematically produce beliefs that mirror reality. Explain why you think you're *rational.* Why you think that, using thought processes like the ones you use, minds will end up believing "snow is white" if and only if snow is white. If you don't believe that the outputs of your thought processes are entangled

with reality, why do you believe the outputs of your thought processes? It's the same thing, or it should be.

4. How Much Evidence Does It Take?[1]

Followup to: What is Evidence?[2]

Previously[3], I defined *evidence* as "an event entangled, by links of cause and effect, with whatever you want to know about", and *entangled* as "happening differently for different possible states of the target". So how much entanglement—how much evidence—is required to support a belief?

Let's start with a question simple enough to be mathematical: how hard would you have to entangle yourself with the lottery[4] in order to win? Suppose there are seventy balls, drawn without replacement, and six numbers to match for the win. Then there are 131,115,985 possible winning combinations, hence a randomly selected ticket would have a 1/131,115,985 probability of winning (0.0000007%). To win the lottery, you would need evidence *selective* enough to visibly favor one combination over 131,115,984 alternatives.

Suppose there are some tests you can perform which discriminate, probabilistically, between winning and losing lottery numbers. For example, you can punch a combination into a little black box that always beeps if the combination is the winner, and has only a 1/4 (25%) chance of beeping if the combination is wrong. In Bayesian[5] terms, we would say the *likelihood ratio* is 4 to 1. This means that the box is 4 times as likely to beep when we punch in a correct combination, compared to how likely it is to beep for an incorrect combination.

There are still a whole lot of possible combinations. If you punch in 20 incorrect combinations, the box will beep on 5 of them by sheer chance (on average). If you punch in all 131,115,985 possible combinations, then while the box is certain

1. *http://lesswrong.com/lw/jn/how_much_evidence_does_it_take/*
2. Page 18, 'What is Evidence?'.
3. Page 18, 'What is Evidence?'.
4. *http://lesswrong.com/lw/hl/lotteries_a_waste_of_hope/*
5. *http://yudkowsky.net/rational/bayes*

to beep for the one winning combination, it will also beep for 32,778,996 losing combinations (on average).

So this box doesn't let you win the lottery, but it's better than nothing. If you used the box, your odds of winning would go from 1 in 131,115,985 to 1 in 32,778,997. You've made some progress toward finding your target, the truth, within the huge space of possibilities.

Suppose you can use another black box to test combinations *twice, independently*. Both boxes are certain to beep for the winning ticket. But the chance of a box beeping for a losing combination is 1/4 *independently* for each box; hence the chance of *both* boxes beeping for a losing combination is 1/16. We can say that the *cumulative* evidence, of two independent tests, has a likelihood ratio of 16:1. The number of losing lottery tickets that pass both tests will be (on average) 8,194,749.

Since there are 131,115,985 possible lottery tickets, you might guess that you need evidence whose strength is around 131,115,985 to 1—an event, or series of events, which is 131,115,985 times more likely to happen for a winning combination than a losing combination. Actually, this amount of evidence would only be enough to give you an *even* chance of winning the lottery. Why? Because if you apply a filter of that power to 131 million losing tickets, there will be, on average, one losing ticket that passes the filter. The winning ticket will also pass the filter. So you'll be left with two tickets that passed the filter, only one of them a winner. 50% odds of winning, if you can only buy one ticket.

A better way of viewing the problem: In the beginning, there is 1 winning ticket and 131,115,984 losing tickets, so your odds of winning are 1:131,115,984. If you use a single box, the odds of it beeping are 1 for a winning ticket and 0.25 for a losing ticket. So we multiply 1:131,115,984 by 1:0.25 and get 1:32,778,996. Adding another box of evidence multiplies the odds by 1:0.25 again, so now the odds are 1 winning ticket to 8,194,749 losing tickets.

It is convenient to measure evidence in bits—not like bits on a hard drive, but mathematician's bits, which are conceptually different. Mathematician's bits are the logarithms, base 1/2, of probabilities. For example, if there are four possible outcomes A, B, C, and D, whose probabilities are 50%, 25%, 12.5%, and 12.5%, and I tell you the outcome was "D", then I have transmitted three bits of information to you, because I informed you of an outcome whose probability was 1/8.

It so happens that 131,115,984 is slightly less than 2 to the 27th power. So 14 boxes or 28 bits of evidence—an event 268,435,456:1 times more likely to happen if the ticket-hypothesis is true than if it is false—would shift the odds from 1:131,115,984 to 268,435,456:131,115,984, which reduces to 2:1. Odds of 2 to 1 mean two chances to win for each chance to lose, so the *probability* of winning with 28 bits of evidence is 2/3. Adding another box, another 2 bits of evidence, would take the odds to 8:1. Adding yet another two boxes would take the chance of winning to 128:1.

So if you want to license a *strong belief* that you will win the lottery—arbitrarily defined as less than a 1% probability of being wrong—34 bits of evidence about the winning combination should do the trick.

In general, the rules for weighing "how much evidence it takes" follow a similar pattern: The larger the *space of possibilities* in which the hypothesis lies, or the more unlikely the hypothesis seems *a priori* compared to its neighbors, or the more confident you wish to be, the more evidence you need.

You cannot defy the rules; you cannot form accurate beliefs based on inadequate evidence. Let's say you've got 10 boxes lined up in a row, and you start punching combinations into the boxes. You cannot stop on the first combination that gets beeps from all 10 boxes, saying, "But the odds of that happening for a losing combination are a million to one! I'll just ignore those ivory-tower Bayesian rules and stop here." On average, 131 losing tickets will pass such a test for every winner. Consid-

ering the space of possibilities and the prior improbability, you jumped to a too-strong conclusion based on insufficient evidence. That's not a pointless bureaucratic regulation, it's math.

Of course, you can still *believe* based on inadequate evidence, if that is your whim; but you will not be able to believe *accurately*. It is like trying to drive your car without any fuel, because you don't believe in the silly-dilly fuddy-duddy concept that it ought to take fuel to go places. It would be so much more *fun,* and so much less expensive, if we just decided to repeal the law that cars need fuel. Isn't it just obviously better for everyone? Well, you can try, if that is your whim. You can even shut your eyes and pretend the car is moving. But to *really* arrive at accurate beliefs requires evidence-fuel, and the further you want to go, the more fuel you need.

5. How to Convince Me That 2 + 2 = 3[1]

In "What is Evidence?", I wrote[2]:

> This is why rationalists put such a heavy premium on the paradoxical-seeming claim that a belief is only really *worthwhile* if you could, in principle, be persuaded to believe otherwise. If your retina ended up in the same state regardless of what light entered it, you would be blind... Hence the phrase, "blind faith". If what you believe doesn't depend on what you see, you've been blinded as effectively as by poking out your eyeballs.

Cihan Baran replied[3]:

> I can not conceive of a situation that would make 2+2 = 4 false. Perhaps for that reason, my belief in 2+2=4 is unconditional.

I admit, I cannot conceive of a "situation" that would *make* 2 + 2 = 4 false. (There are redefinitions, but those are not "situations", and then you're no longer talking about 2, 4, =, or +.) But that doesn't make my belief unconditional. I find it quite easy to imagine a situation which would *convince* me that 2 + 2 = 3.

Suppose I got up one morning, and took out two earplugs, and set them down next to two other earplugs on my nighttable, and noticed that there were now three earplugs, without any earplugs having appeared or disappeared—in contrast to my stored memory that 2 + 2 was supposed to equal 4. Moreover, when I visualized the process in my own mind, it seemed that making XX and XX come out to XXXX required an extra X

1. *http://lesswrong.com/lw/jr/how_to_convince_me_that_2_2_3/*
2. Page 18, 'What is Evidence?'.
3. *http://lesswrong.com/lw/jl/what_is_evidence/f7h*

to appear from nowhere, and was, moreover, inconsistent with other arithmetic I visualized, since subtracting XX from XXX left XX, but subtracting XX from XXXX left XXX. This would conflict with my stored memory that 3 - 2 = 1, but memory would be absurd in the face of physical and mental confirmation that XXX - XX = XX.

I would also check a pocket calculator, Google, and perhaps my copy of 1984 where Winston writes that "Freedom is the freedom to say two plus two equals three." All of these would naturally show that the rest of the world agreed with my current visualization, and disagreed with my memory, that 2 + 2 = 3.

How could I possibly have ever been so deluded as to believe that 2 + 2 = 4? Two explanations would come to mind: First, a neurological fault (possibly caused by a sneeze) had made all the additive sums in my stored memory go up by one. Second, someone was messing with me, by hypnosis or by my being a computer simulation. In the second case, I would think it more likely that they had messed with my arithmetic *recall* than that 2 + 2 *actually* equalled 4. Neither of these plausible-sounding explanations would prevent me from noticing that I was very, very, *very* confused[4].

What would convince me that 2 + 2 = 3, in other words, is exactly the same kind of evidence that currently convinces me that 2 + 2 = 4: The evidential crossfire of physical observation, mental visualization, and social agreement.

There was a time when I had no idea that 2 + 2 = 4. I did not arrive at this *new* belief by random processes—then there would have been no particular reason for my brain to end up storing "2 + 2 = 4" instead of "2 + 2 = 7". The fact that my brain stores an answer surprisingly similar to what happens when I lay down two earplugs alongside two earplugs, calls forth an explanation of what entanglement produces this strange mirroring of mind and reality.

4. Page 62, 'Your Strength as a Rationalist'.

There's really only two possibilities, for a belief of fact[5]—either the belief got there via a mind-reality entangling process[6], or not. If not, the belief can't be correct except by coincidence. For beliefs with the slightest shred of internal complexity[7] (requiring a computer program of more than 10 bits to simulate), the space of possibilities is large enough that coincidence vanishes.

Unconditional facts are not the same as unconditional beliefs. If entangled evidence convinces me that a fact is unconditional, this doesn't mean I always believed in the fact without need of entangled evidence.

I believe that 2 + 2 = 4, and I find it quite easy to conceive of a situation which would convince me that 2 + 2 = 3. Namely, the same sort of situation that currently convinces me that 2 + 2 = 4. Thus I do not fear that I am a victim of blind faith.

If there are any Christians in the audience *who know Bayes's Theorem* (no numerophobes, please) might I inquire of you what situation would convince you of the truth of Islam? Presumably it would be the same sort of situation causally responsible for producing your current belief in Christianity: We would push you screaming out of the uterus of a Muslim woman, and have you raised by Muslim parents who continually told you that it is good to believe unconditionally in Islam. Or is there more to it than that? If so, what situation would convince you of Islam, or at least, non-Christianity?

5. Page 463, 'Feeling Rational'.
6. Page 18, 'What is Evidence?'.
7. Page 29, 'Occam's Razor'.

6. Occam's Razor[1]

Followup to: Burdensome Details[2], How Much Evidence?[3]

The more complex an explanation is, the more evidence you need just to find it in belief-space. (In Traditional Rationality this is often phrased misleadingly[4], as "The more complex a proposition is, the more evidence is required to argue for it.") How can we measure the complexity of an explanation? How can we determine how much evidence is required?

Occam's Razor is often phrased as "The simplest explanation that fits the facts." Robert Heinlein replied that the simplest explanation is "The lady down the street is a witch; she did it."

One observes that the length of an English sentence is not a good way to measure "complexity". And "fitting" the facts by merely *failing to prohibit* them is insufficient.

Why, exactly, is the length of an English sentence a poor measure of complexity? Because when you speak a sentence aloud, you are using *labels* for concepts that the listener shares—the receiver has already stored the complexity in them. Suppose we abbreviated Heinlein's whole sentence as "Tldtsiawsdi!" so that the entire explanation can be conveyed in one word; better yet, we'll give it a short arbitrary label like "Fnord!" Does this reduce the complexity? No, because you have to tell the listener in advance that "Tldtsiawsdi!" stands for "The lady down the street is a witch; she did it." "Witch", itself, is a label for some extraordinary assertions—just because we all know what it means doesn't mean the concept is simple.

An enormous bolt of electricity comes out of the sky and hits something, and the Norse tribesfolk say, "Maybe a really powerful agent was angry and threw a lightning bolt." The hu-

1. *http://lesswrong.com/lw/jp/occams_razor/*
2. *http://lesswrong.com/lw/jk/burdensome_details/*
3. Page 22, 'How Much Evidence Does It Take?'.
4. *http://lesswrong.com/lw/jo/einsteins_arrogance/*

man brain is the most complex artifact in the known universe. If *anger* seems simple, it's because we don't see all the neural circuitry that's implementing the emotion. (Imagine trying to explain why *Saturday Night Live* is funny, to an alien species with no sense of humor. But don't feel superior; you yourself have no sense of fnord.) The complexity of anger, and indeed the complexity of intelligence, was glossed over by the humans who hypothesized Thor the thunder-agent.

To a human, Maxwell's Equations take much longer to explain than Thor. Humans don't have a built-in vocabulary for calculus the way we have a built-in vocabulary for anger. You've got to explain your language, and the language behind the language, and the very concept of mathematics, before you can start on electricity.

And yet it seems that there should be some sense in which Maxwell's Equations are *simpler* than a human brain, or Thor the thunder-agent.

There is: It's *enormously* easier (as it turns out) to write a computer program that simulates Maxwell's Equations, compared to a computer program that simulates an intelligent emotional mind like Thor.

The formalism of Solomonoff Induction measures the "complexity of a description" by the length of the shortest computer program which produces that description as an output. To talk about the "shortest computer program" that does something, you need to specify a space of computer programs, which requires a language and interpreter. Solomonoff Induction uses Turing machines, or rather, bitstrings that specify Turing machines. What if you don't like Turing machines? Then there's only a constant complexity penalty to design your own Universal Turing Machine that interprets whatever code you give it in whatever programming language you like. Different inductive formalisms are penalized by a worst-case constant factor relative to each other, corresponding to the size of a universal interpreter for that formalism.

In the better (IMHO) versions of Solomonoff Induction, the computer program does not produce a deterministic prediction, but assigns probabilities to strings. For example, we could write a program to explain a fair coin by writing a program that assigns equal probabilities to all 2^N strings of length N. This is Solomonoff Induction's approach to *fitting* the observed data. The higher the probability a program assigns to the observed data, the better that program *fits* the data. And probabilities must sum to 1, so for a program to better "fit" one possibility, it must steal probability mass from some other possibility which will then "fit" much more poorly. There is no superfair coin that assigns 100% probability to heads and 100% probability to tails.

How do we trade off the fit to the data, against the complexity of the program? If you ignore complexity penalties, and think *only* about fit, then you will always prefer programs that claim to deterministically predict the data, assign it 100% probability. If the coin shows "HTTHHT", then the program which claims that the coin was fixed to show "HTTHHT" fits the observed data 64 times better than the program which claims the coin is fair. Conversely, if you ignore fit, and consider *only* complexity, then the "fair coin" hypothesis will always seem simpler than any other hypothesis. Even if the coin turns up "HTHHTHHHTHHHHTHHHHHT..." Indeed, the fair coin *is* simpler and it fits this data exactly as well as it fits any other string of 20 coinflips—no more, no less—but we see another hypothesis, seeming not too complicated, that fits the data much better.

If you let a program store one more binary bit of information, it will be able to cut down a space of possibilities by half, and hence assign twice as much probability to all the points in the remaining space. This suggests that one bit of program complexity should cost *at least* a "factor of two gain" in the fit. If you try to design a computer program that explicitly stores an outcome like "HTTHHT", the six bits that you lose in complexity must destroy all plausibility gained by a 64-fold im-

provement in fit. Otherwise, you will sooner or later decide that all fair coins are fixed.

Unless your program is being smart, and *compressing* the data, it should do no good just to move one bit from the data into the program description.

The way Solomonoff induction works to predict sequences is that you sum up over all allowed computer programs—if any program is allowed, Solomonoff induction becomes uncomputable—with each program having a prior probability of (1/2) to the power of its code length in bits, and each program is further weighted by its fit to all data observed so far. This gives you a weighted mixture of experts that can predict future bits.

The Minimum Message Length formalism is nearly equivalent to Solomonoff induction. You send a string describing a code, and then you send a string describing the data in that code. Whichever explanation leads to the shortest *total* message is the best. If you think of the set of allowable codes as a space of computer programs, and the code description language as a universal machine, then Minimum Message Length is nearly equivalent to Solomonoff induction. (Nearly, because it chooses the *shortest* program, rather than summing up over all programs.)

This lets us see clearly the problem with using "The lady down the street is a witch; she did it" to explain the pattern in the sequence "0101010101". If you're sending a message to a friend, trying to describe the sequence you observed, you would have to say: "The lady down the street is a witch; she made the sequence come out 0101010101." Your accusation of witchcraft wouldn't let you *shorten* the rest of the message; you would still have to describe, in full detail, the data which her witchery caused.

Witchcraft may fit our observations in the sense of qualitatively *permitting* them; but this is because witchcraft permits *everything*, like saying "Phlogiston![5]" So, even after you say

5. Page 87, 'Fake Causality'.

"witch", you still have to describe all the observed data in full detail. You have not *compressed the total length of the message describing your observations* by transmitting the message about witchcraft; you have simply added a useless prologue, increasing the total length.

The real sneakiness was concealed in the word "it" of "A witch did it". A witch did *what?*

Of course, thanks to hindsight bias[6] and anchoring[7] and fake explanations[8] and fake causality[9] and positive bias[10] and motivated cognition[11], it may seem all too obvious that if a woman is a witch, of *course* she would make the coin come up 0101010101. But of this I have already spoken.

6. Page 71, 'Hindsight bias'.
7. Page 289, 'Anchoring and Adjustment'.
8. Page 77, 'Fake Explanations'.
9. Page 87, 'Fake Causality'.
10. Page 108, 'Positive Bias: Look Into the Dark'.
11. Page 333, 'Knowing About Biases Can Hurt People'.

7. The Lens That Sees Its Flaws[1]

Continuation of: What is Evidence?[2]

Light leaves the Sun and strikes your shoelaces and bounces off; some photons enter the pupils of your eyes and strike your retina; the energy of the photons triggers neural impulses; the neural impulses are transmitted to the visual-processing areas of the brain; and there the optical information is processed and reconstructed into a 3D model that is recognized as an untied shoelace; and so you believe that your shoelaces are untied.

Here is the secret of *deliberate rationality*—this whole entanglement process is not magic[3], and you can *understand* it. You can *understand* how you see your shoelaces. You can *think* about which sort of thinking processes will create beliefs which mirror reality, and which thinking processes will not.

Mice can see, but they can't understand seeing. *You* can understand seeing, and because of that, you can do things which mice cannot do. Take a moment to marvel[4] at this, for it is indeed marvelous.

Mice see, but they don't know they have visual cortexes, so they can't correct for optical illusions. A mouse lives in a mental world that includes cats, holes, cheese and mousetraps—but not mouse brains. Their camera does not take pictures of its own lens. But we, as humans, can look at a seemingly bizarre image[5], and realize that part of what we're seeing is the lens itself. You don't always have to believe your own eyes, but you have to realize that you *have* eyes—you must have distinct mental buckets for the map and the territory, for the senses and reality. Lest you think this a trivial ability, remember how rare it is in the animal kingdom.

1. *http://lesswrong.com/lw/jm/the_lens_that_sees_its_flaws/*
2. Page 18, 'What is Evidence?'.
3. Page 96, 'Mysterious Answers to Mysterious Questions'.
4. Page 124, '"Science" as Curiosity-Stopper'.
5. *http://www.richrock.com/gifs/optical-illusion-wheels-circles-rotating.png*

The whole idea of Science is, simply, reflective reasoning about a more reliable process for making the contents of your mind mirror the contents of the world. It is the sort of thing mice would never invent. Pondering this business of "performing replicable experiments to falsify theories", we can see *why* it works. Science is not a separate magisterium[6], far away from real life and the understanding of ordinary mortals. Science is not something that only applies to the inside of laboratories[7]. Science, itself, is an understandable process-in-the-world that correlates brains with reality.

Science *makes sense,* when you think about it. But mice can't think about thinking, which is why they don't have Science. One should not overlook the wonder of this—or the potential power it bestows on us as individuals, not just scientific societies.

Admittedly, understanding the engine of thought may be *a little more complicated* than understanding a steam engine—but it is not a *fundamentally* different task.

Once upon a time, I went to EFNet's #philosophy to ask "Do you believe a nuclear war will occur in the next 20 years? If no, why not?" One person who answered the question said he didn't expect a nuclear war for 100 years, because "All of the players involved in decisions regarding nuclear war are not interested right now." "But why extend that out for 100 years?", I asked. "Pure hope," was his reply.

Reflecting on this whole thought process, we can see why the thought of nuclear war makes the person unhappy, and we can see how his brain therefore rejects the belief. But, if you imagine a billion worlds—Everett branches, or Tegmark duplicates[8]—this thought process will not systematically correlate[9] optimists to branches in which no nuclear war occurs. (Some

6. *http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/*
7. *http://lesswrong.com/lw/gv/outside_the_laboratory/*
8. *http://arxiv.org/abs/astro-ph/0302131*
9. Page 18, 'What is Evidence?'.

clever fellow is bound to say, "Ah, but since I have hope, I'll work a little harder at my job, pump up the global economy, and thus help to prevent countries from sliding into the angry and hopeless state where nuclear war is a possibility. So the two events are related after all." At this point, we have to drag in Bayes's Theorem[10] and measure the charge of entanglement quantitatively. Your optimistic nature cannot have *that* large an effect on the world; it cannot, of itself, decrease the probability of nuclear war by 20%, or however much your optimistic nature shifted your beliefs. Shifting your beliefs by a large amount, due to an event that only carries a very tiny charge of entanglement, will still mess up your mapping.)

To ask which beliefs make you happy, is to turn inward, not outward—it tells you something about yourself, but it is not evidence entangled with the environment. I have nothing anything against happiness, but it should follow from[11] your picture of the world, rather than tampering with the mental paintbrushes.

If you can see this—if you can see that hope is shifting your *first-order* thoughts by too large a degree—if you can understand your mind as a mapping-engine with flaws in it—then you can apply a reflective correction. The brain is a flawed lens through which to see reality. This is true of both mouse brains and human brains. But a human brain is a flawed lens that can understand its own flaws—its systematic errors, its biases—and apply second-order corrections to them. This, *in practice,* makes the flawed lens far more powerful. Not perfect, but far more powerful.

10. *http://yudkowsky.net/rational/bayes*
11. Page 463, 'Feeling Rational'.

Part II
Mysterious Answers to Mysterious Questions

A sequence on how to see through the disguises of answers or beliefs or statements, that don't answer or say or mean anything.

1. Making Beliefs Pay Rent (in Anticipated Experiences)[1]

Thus begins the ancient parable:

If a tree falls in a forest and no one hears it, does it make a sound? One says, "Yes it does, for it makes vibrations in the air." Another says, "No it does not, for there is no auditory processing in any brain."

Suppose that, after the tree falls, the two walk into the forest together. Will one expect to see the tree fallen to the right, and the other expect to see the tree fallen to the left? Suppose that before the tree falls, the two leave a sound recorder next to the tree. Would one, playing back the recorder, expect to hear something different from the other? Suppose they attach an electroencephalograph to any brain in the world; would one expect to see a different trace than the other? Though the two argue, one saying "No," and the other saying "Yes," they do not anticipate any different experiences. The two think they have different models of the world, but they have no difference with respect to what they expect will *happen to* them.

It's tempting to try to eliminate this mistake class by insisting that the only legitimate kind of belief is an anticipation of sensory experience. But the world does, in fact, contain much that is not sensed directly. We don't see the atoms underlying the brick, but the atoms are in fact there. There is a floor beneath your feet, but you don't *experience* the floor directly; you see the light *reflected* from the floor, or rather, you see what your retina and visual cortex have processed of that light. To infer the floor from seeing the floor is to step back into the unseen causes of experience. It may seem like a very short and direct step, but it is still a step.

You stand on top of a tall building, next to a grandfather clock with an hour, minute, and ticking second hand. In your

1. *http://lesswrong.com/lw/i3/making_beliefs_pay_rent_in_anticipated_experiences/*

hand is a bowling ball, and you drop it off the roof. On which tick of the clock will you hear the crash of the bowling ball hitting the ground?

To answer precisely, you must use beliefs like *Earth's gravity is 9.8 meters per second per second,* and *This building is around 120 meters tall.* These beliefs are not wordless anticipations of a sensory experience; they are verbal-ish, propositional. It probably does not exaggerate much to describe these two beliefs as sentences made out of words. But these two beliefs have an inferential *consequence* that is a direct sensory anticipation—if the clock's second hand is on the 12 numeral when you drop the ball, you anticipate seeing it on the 1 numeral when you hear the crash five seconds later. To anticipate sensory experiences as precisely as possible, we must process beliefs that are not anticipations of sensory experience.

It is a great strength of *Homo sapiens* that we can, better than any other species in the world, learn to model the unseen. It is also one of our great weak points. Humans often believe in things that are not only unseen but unreal.

The same brain that builds a network of inferred causes behind sensory experience, can also build a network of causes that is not connected to sensory experience, or poorly connected. Alchemists believed that phlogiston caused fire—we could oversimply their minds by drawing a little node labeled "Phlogiston", and an arrow from this node to their sensory experience of a crackling campfire—but this belief yielded no advance predictions; the link from phlogiston to experience was always configured after the experience, rather than constraining the experience in advance. Or suppose your postmodern English professor teaches you that the famous writer Wulky Wilkinsen is actually a "post-utopian". What does this mean you should expect from his books? Nothing. The belief, if you can call it that, doesn't connect to sensory experience at all. But you had better remember the propositional assertion that "Wulky Wilkinsen" has the "post-utopian" attribute, so you can regurgitate it on

the upcoming quiz. Likewise if "post-utopians" show "colonial alienation"; if the quiz asks whether Wulky Wilkinsen shows colonial alienation, you'd better answer yes. The beliefs are connected to each other, though still not connected to any anticipated experience.

We can build up whole networks of beliefs that are connected only to each other—call these "floating" beliefs. It is a uniquely human flaw among animal species, a perversion of *Homo sapiens's* ability to build more general and flexible belief networks.

The rationalist virtue of *empiricism* consists of constantly asking which experiences our beliefs predict—or better yet, prohibit. Do you believe that phlogiston is the cause of fire? Then what do you expect to see happen, because of that? Do you believe that Wulky Wilkinsen is a post-utopian? Then what do you expect to see because of that? No, not "colonial alienation"; *what experience will happen to you?* Do you believe that if a tree falls in the forest, and no one hears it, it still makes a sound? Then what experience must therefore befall you?

It is even better to ask: what experience *must not* happen to you? Do you believe that *elan vital* explains the mysterious aliveness of living beings? Then what does this belief *not* allow to happen—what would definitely falsify this belief? A null answer means that your belief does not *constrain* experience; it permits *anything* to happen to you. It floats.

When you argue a seemingly factual question, always keep in mind which difference of anticipation you are arguing about. If you can't find the difference of anticipation, you're probably arguing about labels in your belief network—or even worse, floating beliefs, barnacles on your network. If you don't know what experiences are implied by Wulky Wilkinsen being a post-utopian, you can go on arguing forever. (You can also publish papers forever.)

Above all, don't ask what to believe—ask what to anticipate. Every question of belief should flow from a question of antici-

pation, and that question of anticipation should be the center of the inquiry. Every guess of belief should begin by flowing to a specific guess of anticipation, and should continue to pay rent in future anticipations. If a belief turns deadbeat, evict it.

2. Belief in Belief[1]

Followup to: Making Beliefs Pay Rent (in Anticipated Experiences)[2]

Carl Sagan once told a parable[3] of a man who comes to us and claims: "There is a dragon in my garage." Fascinating! We reply that we wish to see this dragon—let us set out at once for the garage! "But wait," the claimant says to us, "it is an *invisible* dragon."

Now as Sagan points out, this doesn't make the hypothesis unfalsifiable. Perhaps we go to the claimant's garage, and although we see no dragon, we hear heavy breathing from no visible source; footprints mysteriously appear on the ground; and instruments show that something in the garage is consuming oxygen and breathing out carbon dioxide.

But now suppose that we say to the claimant, "Okay, we'll visit the garage and see if we can hear heavy breathing," and the claimant quickly says no, it's an *inaudible* dragon. We propose to measure carbon dioxide in the air, and the claimant says the dragon does not breathe. We propose to toss a bag of flour into the air to see if it outlines an invisible dragon, and the claimant immediately says, "The dragon is permeable to flour."

Carl Sagan used this parable to illustrate the classic moral that poor hypotheses need to do fast footwork to avoid falsification. But I tell this parable to make a different point: The claimant must have an accurate model of the situation *somewhere* in his mind, because he can anticipate, in advance, *exactly which experimental results he'll need to excuse.*

Some philosophers have been much confused by such scenarios, asking, "Does the claimant *really* believe there's a dragon present, or not?" As if the human brain only had enough disk space to represent one belief at a time! Real minds are

1. *http://lesswrong.com/lw/i4/belief_in_belief/*
2. Page 39, 'Making Beliefs Pay Rent (in Anticipated Experiences)'.
3. *http://www.godlessgeeks.com/LINKS/Dragon.htm*

more tangled than that. As discussed in yesterday's post, there are different types of belief; not all beliefs are direct anticipations[4]. The claimant clearly does not *anticipate* seeing anything unusual upon opening the garage door; otherwise he wouldn't make advance excuses. It may also be that the claimant's pool of propositional beliefs contains *There is a dragon in my garage.* It may seem, to a rationalist, that these two beliefs should collide and conflict even though they are of different types. Yet it is a physical fact that you can write "The sky is green!" next to a picture of a blue sky without the paper bursting into flames.

The rationalist virtue of empiricism is supposed to prevent us from this class of mistake. We're supposed to constantly ask our beliefs which experiences they predict, make them pay rent in anticipation. But the dragon-claimant's problem runs deeper, and cannot be cured with such simple advice. It's not exactly *difficult* to connect belief in a dragon to anticipated experience of the garage. If you believe there's a dragon in your garage, then you can expect to open up the door and see a dragon. If you don't see a dragon, then that means there's no dragon in your garage. This is pretty straightforward. You can even try it with your own garage.

No, this invisibility business is a symptom of something much worse.

Depending on how your childhood went, you may remember a time period when you first began to doubt Santa Claus's existence, but you still believed that you were *supposed* to believe in Santa Claus, so you tried to deny the doubts. As Daniel Dennett observes, where it is difficult to believe a thing, it is often much easier to believe that you *ought* to believe it. What does it mean to believe that the Ultimate Cosmic Sky[5] is both perfectly blue and perfectly green? The statement is confusing; it's not even clear what it would *mean* to believe it—what exactly would *be* believed, if you believed. You can much more easily

4. Page 39, 'Making Beliefs Pay Rent (in Anticipated Experiences)'.
5. Page 143, 'A Fable of Science and Politics'.

believe that it is *proper,* that it is *good* and *virtuous* and *beneficial,* to believe that the Ultimate Cosmic Sky is both perfectly blue and perfectly green. Dennett calls this "belief in belief".

And here things become complicated, as human minds are wont to do—I think even Dennett oversimplifies how this psychology works in practice. For one thing, if you believe in belief, you cannot admit to yourself that you only believe in belief, because it is virtuous to *believe,* not to believe in belief, and so if you only believe in belief, instead of believing, you are not virtuous. Nobody will *admit* to themselves, "I don't believe the Ultimate Cosmic Sky is blue and green, but I believe I ought to believe it"—not unless they are unusually capable of acknowledging their own lack of virtue. People don't believe in belief in belief, they just believe in belief.

(Those who find this confusing may find it helpful to study mathematical logic, which trains one to make very sharp distinctions between the proposition P, a proof of P, and a proof that P is provable. There are similarly sharp distinctions between P, wanting P, believing P, wanting to believe P, and believing that you believe P.)

There's different kinds of belief in belief. You may believe in belief explicitly; you may recite in your deliberate stream of consciousness the verbal sentence "It is virtuous to believe that the Ultimate Cosmic Sky is perfectly blue and perfectly green." (While also believing that you believe this, unless you are unusually capable of acknowledging your own lack of virtue.) But there's also less explicit forms of belief in belief. Maybe the dragon-claimant fears the public ridicule that he imagines will result if he publicly confesses he was wrong (although, in fact, a rationalist would congratulate him, and others are more likely to ridicule him if he goes on claiming there's a dragon in his garage). Maybe the dragon-claimant flinches away from the prospect of admitting to himself that there is no dragon, because it conflicts with his self-image as the glorious discoverer

of the dragon, who saw in his garage what all others had failed to see.

If all our thoughts were deliberate verbal sentences like philosophers manipulate, the human mind would be a great deal easier for humans to understand. Fleeting mental images, unspoken flinches, desires acted upon without acknowledgement—these account for as much of ourselves as words.

While I disagree with Dennett on some details and complications, I still think that Dennett's notion of *belief in belief* is the key insight necessary to understand the dragon-claimant. But we need a wider concept of *belief,* not limited to verbal sentences. "Belief" should include unspoken anticipation-controllers. "Belief in belief" should include unspoken cognitive-behavior-guiders. It is not psychologically realistic to say "The dragon-claimant does not believe there is a dragon in his garage; he believes it is beneficial to believe there is a dragon in his garage." But it is realistic to say the dragon-claimant *anticipates as if* there is no dragon in his garage, and *makes excuses as if* he believed in the belief.

You can possess an ordinary mental picture of your garage, with no dragons in it, which correctly predicts your experiences on opening the door, and never once think the verbal phrase *There is no dragon in my garage*. I even bet it's happened to you—that when you open your garage door or bedroom door or whatever, and expect to see no dragons, no such verbal phrase runs through your mind.

And to flinch away from giving up your belief in the dragon—or flinch away from giving up your *self-image* as a person who believes in the dragon—it is not necessary to explicitly think *I want to believe there's a dragon in my garage*. It is only necessary to flinch away from the prospect of admitting you don't believe.

To correctly anticipate, in advance, which experimental results shall need to be excused, the dragon-claimant must (a) possess an accurate anticipation-controlling model somewhere

in his mind, and (b) act cognitively to protect either (b1) his free-floating propositional belief in the dragon or (b2) his self-image of believing in the dragon.

If someone believes in their belief in the dragon, and also believes in the dragon, the problem is much less severe. They will be willing to stick their neck out on experimental predictions, and perhaps even agree to give up the belief if the experimental prediction is wrong—although belief in belief can still interfere with this, if the belief itself is not absolutely confident. When someone makes up excuses *in advance,* it would seem to require that belief, and belief in belief, have become unsynchronized.

3. Bayesian Judo[1]

You can have some fun with people whose anticipations get out of sync with what they believe they believe[2].

I was once at a dinner party, trying to explain to a man what I did for a living, when he said: "I don't believe Artificial Intelligence is possible because only God can make a soul."

At this point I must have been divinely inspired, because I instantly responded: "You mean if I can make an Artificial Intelligence, it proves your religion is false?"

He said, "What?"

I said, "Well, if your religion predicts that I can't possibly make an Artificial Intelligence, then, if I make an Artificial Intelligence, it means your religion is false. Either your religion allows that it might be possible for me to build an AI; or, if I build an AI, that disproves your religion."

There was a pause, as the one realized he had just made his hypothesis vulnerable to falsification, and then he said, "Well, I didn't mean that you couldn't make an intelligence, just that it couldn't be emotional in the same way we are."

I said, "So if I make an Artificial Intelligence that, without being deliberately preprogrammed with any sort of script, starts talking about an emotional life that sounds like ours, *that* means your religion is wrong."

He said, "Well, um, I guess we may have to agree to disagree on this."

I said: "No, we can't, actually. There's a theorem of rationality called Aumann's Agreement Theorem which shows that no two rationalists can agree to disagree. If two people disagree with each other, at least one of them must be doing something wrong."

1. *http://lesswrong.com/lw/i5/bayesian_judo/*
2. Page 43, 'Belief in Belief'.

We went back and forth on this briefly. Finally, he said, "Well, I guess I was really trying to say that I don't think you can make something eternal."

I said, "Well, I don't think so either! I'm glad we were able to reach agreement on this, as Aumann's Agreement Theorem requires." I stretched out my hand, and he shook it, and then he wandered away.

A woman who had stood nearby, listening to the conversation, said to me gravely, "That was beautiful."

"Thank you very much," I said.

4. Professing and Cheering[1]

I once attended a panel on the topic, "Are science and religion compatible?" One of the women on the panel, a pagan, held forth interminably upon how she believed that the Earth had been created when a giant primordial cow was born into the primordial abyss, who licked a primordial god into existence, whose descendants killed a primordial giant and used its corpse to create the Earth, etc. The tale was long, and detailed, and more absurd than the Earth being supported on the back of a giant turtle. And the speaker clearly knew enough science to know this.

I still find myself struggling for words to describe what I saw as this woman spoke. She spoke with... pride? Self-satisfaction? A deliberate flaunting of herself?

The woman went on describing her creation myth for what seemed like forever, but was probably only five minutes. That strange pride/satisfaction/flaunting clearly had something to do with her *knowing* that her beliefs were scientifically outrageous. And it wasn't that she hated science; as a panelist she professed that religion and science were compatible. She even talked about how it was quite understandable that the Vikings talked about a primordial abyss, given the land in which they lived—explained away her own religion!—and yet nonetheless insisted this was what she "believed", said with peculiar satisfaction.

I'm not sure that Daniel Dennett's concept of "belief in belief[2]" stretches to cover this event. It was weirder than that. She didn't recite her creation myth with the fanatical faith of someone who needs to reassure herself. She didn't act like she expected us, the audience, to be convinced—or like she needed our belief to validate her.

1. *http://lesswrong.com/lw/i6/professing_and_cheering/*
2. Page 43, 'Belief in Belief'.

Dennett, in addition to suggesting belief in belief, has also suggested that much of what is called "religious belief" should really be studied as "religious profession". Suppose an alien anthropologist studied a group of postmodernist English students who all seemingly *believed* that Wulky Wilkensen was a post-utopian author. The appropriate question may not be "Why do the students all believe this strange belief?" but "Why do they all write this strange sentence on quizzes?" Even if a sentence is essentially meaningless, you can still know when you are supposed to chant the response aloud.

I think Dennett may be slightly too cynical in suggesting that religious profession is *just* saying the belief aloud—most people are honest enough that, if they say a religious statement aloud, they will also feel obligated to say the verbal sentence into their own stream of consciousness.

But even the concept of "religious profession" doesn't seem to cover the pagan woman's claim to believe in the primordial cow. If you had to profess a religious belief to satisfy a priest, or satisfy a co-religionist—heck, to satisfy your own self-image as a religious person—you would have to *pretend* to believe *much more convincingly* than this woman was doing. As she recited her tale of the primordial cow, with that same strange flaunting pride, she wasn't even *trying* to be persuasive—wasn't even trying to convince us that she took her own religion seriously. I think that's the part that so took me aback. I know people who believe they believe ridiculous things, but when they profess them, they'll spend much more effort to convince themselves that they take their beliefs seriously.

It finally occurred to me that this woman wasn't trying to convince us or even convince herself. Her recitation of the creation story wasn't *about* the creation of the world at all. Rather, by launching into a five-minute diatribe about the primordial cow, she was *cheering for paganism,* like holding up a banner at a football game. A banner saying "GO BLUES[3]" isn't a state-

3. Page 143, 'A Fable of Science and Politics'.

ment of fact, or an attempt to persuade; it doesn't have to be convincing—it's a cheer.

That strange flaunting pride... it was like she was marching naked in a gay pride parade. (Incidentally, I'd have no objection if she *had* marched naked in a gay pride parade. Lesbianism is not something that truth can destroy[4].) It wasn't just a cheer, like marching, but an outrageous cheer, like marching naked—believing that she couldn't be arrested or criticized, because she was doing it for her pride parade.

That's why it mattered to her that what she was saying was beyond ridiculous. If she'd tried to make it sound more plausible, it would have been like putting on clothes.

4. Page 463, 'Feeling Rational'.

5. Belief as Attire[1]

I have so far distinguished between belief as anticipation-controller[2], belief in belief[3], professing and cheering[4]. Of these, we might call anticipation-controlling beliefs "proper beliefs" and the other forms "improper belief". A proper belief can be wrong or irrational, e.g., someone who genuinely anticipates that prayer will cure her sick baby, but the other forms are arguably "not belief at all".

Yet another form of improper belief is belief as group-identification—as a way of belonging. Robin Hanson uses the excellent metaphor[5] of wearing unusual clothing, a group uniform like a priest's vestments or a Jewish skullcap, and so I will call this "belief as attire".

In terms of humanly realistic psychology[6], the Muslims who flew planes into the World Trade Center undoubtedly saw themselves as heroes defending truth, justice, and the Islamic Way from hideous alien monsters a la the movie Independence Day[7]. Only a very inexperienced nerd, the sort of nerd who has no idea how non-nerds see the world, would say this out loud in an Alabama bar. It is not an American thing to say. The American thing to say is that the terrorists "hate our freedom" and that flying a plane into a building is a "cowardly act". You cannot say the phrases "heroic self-sacrifice" and "suicide bomber" in the same sentence, even for the sake of accurately describing how the Enemy sees the world. The very *concept* of the courage and altruism of a suicide bomber is Enemy attire—you can tell, because the Enemy talks about it. The cowardice and sociopathy of a suicide bomber is American attire. There are no

1. *http://lesswrong.com/lw/i7/belief_as_attire/*
2. Page 39, 'Making Beliefs Pay Rent (in Anticipated Experiences)'.
3. Page 43, 'Belief in Belief'.
4. Page 50, 'Professing and Cheering'.
5. *http://lesswrong.com/lw/i6/professing_and_cheering/egb*
6. Page 160, 'Are Your Enemies Innately Evil?'.
7. *http://www.imdb.com/title/tt0116629/*

quote marks you can use to talk about how the Enemy sees the world; it would be like dressing up as a Nazi for Halloween.

Belief-as-attire may help explain how people can be *passionate* about improper beliefs. Mere belief in belief[8], or religious professing[9], would have some trouble creating genuine, deep, powerful emotional effects. Or so I suspect; I confess I'm not an expert here. But my impression is this: People who've stopped anticipating-as-if their religion is true, will go to great lengths to *convince* themselves they are passionate, and this desperation can be mistaken for passion. But it's not the same fire they had as a child.

On the other hand, it is very easy for a human being to genuinely, passionately, gut-level belong to a group, to cheer for their favorite sports team[10]. (This is the foundation on which rests the swindle of "Republicans vs. Democrats" and analogous false dilemmas[11] in other countries, but that's a topic for another post.) Identifying with a tribe is a very strong emotional force. People will die for it. And once you get people to identify with a tribe, the beliefs which are attire of that tribe will be spoken with the full passion of belonging to that tribe.

8. Page 43, 'Belief in Belief'.
9. Page 50, 'Professing and Cheering'.
10. Page 143, 'A Fable of Science and Politics'.
11. Page 426, 'The Third Alternative'.

6. Focus Your Uncertainty[1]

Will bond yields go up, or down, or remain the same? If you're a TV pundit and your job is to explain the outcome after the fact, then there's no reason to worry. No matter *which* of the three possibilities comes true, you'll be able to explain why the outcome perfectly fits your pet market theory . There's no reason to think of these three possibilities as somehow *opposed* to one another, as *exclusive,* because you'll get full marks for punditry no matter which outcome occurs.

But wait! Suppose you're a *novice* TV pundit, and you aren't experienced enough to make up plausible explanations on the spot. You need to prepare remarks in advance for tomorrow's broadcast, and you have limited time to prepare. In this case, it would be helpful to know *which* outcome will actually occur—whether bond yields will go up, down, or remain the same—because then you would only need to prepare *one* set of excuses.

Alas, no one can possibly foresee the future. What are you to do? You certainly can't use "probabilities". We all know from school[2] that "probabilities" are little numbers that appear next to a word problem, and there aren't any little numbers here. Worse, you *feel* uncertain. You don't remember *feeling* uncertain while you were manipulating the little numbers in word problems. *College classes teaching math* are nice clean places, therefore *math itself* can't apply to life situations that aren't nice and clean. You wouldn't want to inappropriately transfer thinking skills from one context to another[3]. Clearly, this is not a matter for "probabilities".

Nonetheless, you only have 100 minutes to prepare your excuses. You can't spend the entire 100 minutes on "up", and also

1. *http://lesswrong.com/lw/ia/focus_your_uncertainty/*
2. *http://lesswrong.com/lw/i2/two_more_things_to_unlearn_from_school/*
3. *http://www.aft.org/pubs-reports/american_educator/issues/summer07/Crit_Thinking.pdf*

spend all 100 minutes on "down", and also spend all 100 minutes on "same". You've got to prioritize somehow.

If you needed to justify your time expenditure to a review committee, you would have to spend equal time on each possibility. Since there are no little numbers written down, you'd have no documentation to justify spending different amounts of time. You can hear the reviewers now: *And why, Mr. Finkledinger, did you spend exactly 42 minutes on excuse #3? Why not 41 minutes, or 43? Admit it—you're not being objective! You're playing subjective favorites!*

But, you realize with a small flash of relief, there's no review committee to scold you. This is good, because there's a major Federal Reserve announcement tomorrow, and it seems unlikely that bond prices will remain the same. You don't want to spend 33 precious minutes on an excuse you don't anticipate needing.

Your mind keeps drifting to the explanations you use on television, of why each event plausibly fits your market theory. But it rapidly becomes clear that plausibility can't help you here—all three events are plausible. Fittability to your pet market theory doesn't tell you how to divide your time. There's an uncrossable gap between your 100 minutes of time, which are conserved; versus your ability to explain how an outcome fits your theory, which is unlimited.

And yet... even in your uncertain state of mind, it seems that you *anticipate* the three events differently; that you *expect* to need some excuses more than others. And—this is the fascinating part—when you think of something that makes it seem *more* likely that bond prices will go up, then you feel *less* likely to need an excuse for bond prices going down or remaining the same.

It even seems like there's a relation between how much you anticipate each of the three outcomes, and how much time you want to spend preparing each excuse. Of course the relation can't actually be quantified. You have 100 minutes to prepare

your speech, but there isn't 100 of anything to divide up in this anticipation business. (Although you do work out that, *if* some particular outcome occurs, then your utility function is logarithmic in time spent preparing the excuse.)

Still... your mind keeps coming back to the idea that anticipation is limited, unlike excusability, but like time to prepare excuses. Maybe anticipation should be treated as a *conserved resource,* like money. Your first impulse is to try to get more anticipation, but you soon realize that, even if you get more anticiptaion, you won't have any more time to prepare your excuses. No, your only course is to *allocate* your *limited supply* of anticipation as best you can.

You're pretty sure you weren't taught anything like that in your statistics courses. They didn't tell you what to do when you *felt* so terribly uncertain. They didn't tell you what to do when there were no little numbers handed to you. Why, even if you tried to use numbers, you might end up using any sort of numbers at all—there's no hint what kind of math to use, if you should be using math! Maybe you'd end up using *pairs* of numbers, right and left numbers, which you'd call DS for Dexter-Sinister... or who knows what else? (Though you do have only 100 minutes to spend preparing excuses.)

If only there were an art of *focusing your uncertainty*—of *squeezing* as much anticipation as possible into whichever outcome will *actually happen!*

But what could we call an art like that? And what would the rules be like?

7. The Virtue of Narrowness[1]

What is true of one apple may not be true of another apple; thus more can be said about a single apple than about all the apples in the world.

—Twelve Virtues of Rationality[2] [3]

Within their own professions, people grasp the importance of narrowness; a car mechanic knows the difference between a carburetor and a radiator, and would not think of them both as "car parts". A hunter-gatherer knows the difference between a lion and a panther. A janitor does not wipe the floor with window cleaner, even if the bottles look similar to one who has not mastered the art.

Outside their own professions, people often commit the misstep of trying to broaden a word as widely as possible, to cover as much territory as possible. Is it not more glorious, more wise, more impressive, to talk about *all* the apples in the world? How much loftier it must be to *explain human thought in general,* without being distracted by smaller questions, such as how humans invent techniques for solving a Rubik's Cube. Indeed, it scarcely seems necessary to consider *specific* questions at all; isn't a general theory a worthy enough accomplishment on its own?

It is the way of the curious to lift up one pebble from among a million pebbles on the shore, and see something new about it, something interesting, something different. You call these pebbles "diamonds", and ask what might be special about them—what inner qualities they might have in common, beyond the glitter you first noticed. And then someone else comes along and says: "Why not call *this* pebble a diamond too? And this one, and this one?" They are enthusiastic, and they mean well. For it seems undemocratic and exclusionary and elitist

1. *http://lesswrong.com/lw/ic/the_virtue_of_narrowness/*
2. *http://yudkowsky.net/virtues/*
3. *http://yudkowsky.net/virtues/*

and unholistic to call some pebbles "diamonds", and others not. It seems... *narrow-minded*... if you'll pardon the phrase. Hardly *open,* hardly *embracing,* hardly *communal.*

You might think it poetic, to give one word many meanings, and thereby spread shades of connotation all around. But even poets, if they are good poets, must learn to see the world precisely. It is not enough to compare love to a flower. Hot jealous unconsummated love is not the same as the love of a couple married for decades. If you need a flower to symbolize jealous love, you must go into the garden, and look, and make subtle distinctions—find a flower with a heady scent, and a bright color, and thorns. Even if your intent is to shade meanings and cast connotations, you must keep precise track of exactly which meanings you shade and connote.

It is a necessary part of the rationalist's art—or even the poet's art!—to focus narrowly on unusual pebbles which possess some special quality. And look at the details which those pebbles—and those pebbles alone!—share among each other. This is not a sin.

It is perfectly all right for modern evolutionary biologists to explain *just* the patterns of living creatures, and not the "evolution" of stars or the "evolution" of technology. Alas, some unfortunate souls use the same word "evolution" to cover the naturally selected patterns of replicating life, *and* the strictly accidental structure of stars, *and* the intelligently configured structure of technology. And as we all know, if people use the same word, it must all be the same thing. You should automatically generalize anything you think you know about biological evolution to technology. Anyone who tells you otherwise must be a mere pointless pedant. It couldn't possibly be that your abysmal ignorance of modern evolutionary theory is so total that you can't tell the difference between a carburetor and a radiator. That's unthinkable. No, the *other* guy—you know, the one who's studied the math—is just too dumb to see the connections.

And what could be more virtuous than seeing connections? Surely the wisest of all human beings are the New Age gurus who say "Everything is connected to everything else." If you ever say this aloud, you should pause, so that everyone can absorb the sheer shock of this Deep Wisdom.

There is a trivial mapping between a graph and its complement. A fully connected graph, with an edge between every two vertices, conveys the same amount of information as a graph with no edges at all. The important graphs are the ones where some things are *not* connected to some other things.

When the unenlightened ones try to be profound, they draw endless verbal comparisons between this topic, and that topic, which is like this, which is like that; until their graph is fully connected and also totally useless. The remedy is specific knowledge and in-depth study. When you understand things in detail, you can see how they are *not* alike, and start enthusiastically subtracting edges *off* your graph.

Likewise, the important categories are the ones that do not contain everything in the universe. Good hypotheses can only explain some possible outcomes, and not others.

It was perfectly all right for Isaac Newton to explain *just* gravity, *just* the way things fall down—and how planets orbit the Sun, and how the Moon generates the tides—but *not* the role of money in human society or how the heart pumps blood. Sneering at narrowness is rather reminiscent of ancient Greeks who thought that going out and actually *looking* at things was manual labor, and manual labor was for slaves.

As Plato put it (in *The Republic, Book VII*):

> "If anyone should throw back his head and learn something by staring at the varied patterns on a ceiling, apparently you would think that he was contemplating with his reason, when he was only staring with his eyes... I cannot but believe that no study makes the soul look on high except that which

> is concerned with real being and the unseen. Whether he gape and stare upwards, or shut his mouth and stare downwards, if it be things of the senses that he tries to learn something about, I declare he never could learn, for none of these things admit of knowledge: I say his soul is looking down, not up, even if he is floating on his back on land or on sea!"

Many today make a similar mistake, and think that narrow concepts are as lowly and unlofty and unphilosophical as, say, going out and looking at things—an endeavor only suited to the underclass. But rationalists—and also poets—need narrow words to express precise thoughts; they need categories which include only some things, and exclude others. There's nothing wrong with focusing your mind, narrowing your categories, excluding possibilities, and sharpening your propositions. Really, there isn't! If you make your words too broad, you end up with something that isn't true and doesn't even make good poetry.

And DON'T EVEN GET ME STARTED on people who think Wikipedia is an "Artificial Intelligence", the invention of LSD was a "Singularity" or that corporations are "superintelligent"!

8. Your Strength as a Rationalist[1]

(The following happened to me in an IRC chatroom, long enough ago that I was still hanging around in IRC chatrooms. Time has fuzzed the memory and my report may be imprecise.)

So there I was, in an IRC chatroom, when someone reports that a friend of his needs medical advice. His friend says that he's been having sudden chest pains, so he called an ambulance, and the ambulance showed up, but the paramedics told him it was nothing, and left, and now the chest pains are getting worse. What should his friend do?

I was confused by this story. I remembered reading about homeless people in New York who would call ambulances just to be taken someplace warm, and how the paramedics always had to take them to the emergency room, even on the 27th iteration. Because if they didn't, the ambulance company could be sued for lots and lots of money. Likewise, emergency rooms are legally obligated to treat anyone, regardless of ability to pay. (And the hospital absorbs the costs, which are enormous, so hospitals are closing their emergency rooms... It makes you wonder what's the point of having economists if we're just going to ignore them.) So I didn't quite understand how the described events could have happened. *Anyone* reporting sudden chest pains should have been hauled off by an ambulance instantly.

And this is where I fell down as a rationalist. I remembered several occasions where my doctor would completely fail to panic at the report of symptoms that seemed, to me, very alarming. And the Medical Establishment was always right. Every single time. I had chest pains myself, at one point, and the doctor patiently explained to me that I was describing chest muscle pain, not a heart attack. So I said into the IRC channel, "Well, if the paramedics told your friend it was nothing, it must *really*

1. *http://lesswrong.com/lw/if/your_strength_as_a_rationalist/*

be nothing—they'd have hauled him off if there was the tiniest chance of serious trouble."

Thus I managed to explain the story within my existing model, though the fit still felt a little forced…

Later on, the fellow comes back into the IRC chatroom and says his friend made the whole thing up. Evidently this was not one of his more reliable friends.

I should have realized, perhaps, that an unknown acquaintance of an acquaintance in an IRC channel might be less reliable[2] than a published journal article. Alas, belief is easier than disbelief; we believe instinctively, but disbelief requires a conscious effort[3].

So instead, by dint of mighty straining, I forced my model of reality to explain an anomaly that *never actually happened*. And I *knew* how embarrassing this was. I *knew* that the usefulness of a model is not what it can explain, but what it can't. A hypothesis that forbids nothing, permits everything, and thereby fails to constrain anticipation[4].

Your strength as a rationalist is your ability to be more confused by fiction than by reality. If you are equally good at explaining any outcome, you have zero knowledge.

We are all weak, from time to time; the sad part is that I *could* have been stronger. I had all the information I needed to arrive at the correct answer, I even *noticed* the problem, and then I ignored it. My feeling of confusion was a Clue, and I threw my Clue away.

I should have paid more attention to that sensation of *still feels a little forced*. It's one of the most important feelings a truthseeker can have, a part of your strength as a rationalist. It is a design flaw in human cognition that this sensation manifests as a quiet strain in the back of your mind, instead of a

2. *http://www.overcomingbias.com/2007/08/truth-bias.html*
3. *http://www.wjh.harvard.edu/%7Edtg/Gilbert%20et%20al%20%28EVERYTHING%20YOU%20READ%29.pdf*
4. Page 39, 'Making Beliefs Pay Rent (in Anticipated Experiences)'.

wailing alarm siren and a glowing neon sign reading "EITHER YOUR MODEL IS FALSE OR THIS STORY IS WRONG."

9. Absence of Evidence Is Evidence of Absence[1]

From Robyn Dawes's *Rational Choice in an Uncertain World*:

> Post-hoc fitting of evidence to hypothesis was involved in a most grievous chapter in United States history: the internment of Japanese-Americans at the beginning of the Second World War. When California governor Earl Warren testified before a congressional hearing in San Francisco on February 21, 1942, a questioner pointed out that there had been no sabotage or any other type of espionage by the Japanese-Americans up to that time. Warren responded, "I take the view that this lack [of subversive activity] is the most ominous sign in our whole situation. It convinces me more than perhaps any other factor that the sabotage we are to get, the Fifth Column activities are to get, are timed just like Pearl Harbor was timed... I believe we are just being lulled into a false sense of security."

Consider Warren's argument from a Bayesian perspective[2]. When we see evidence, hypotheses that assigned a *higher* likelihood to that evidence, gain probability at the expense of hypotheses that assigned a *lower* likelihood to the evidence. This is a phenomenon of *relative* likelihoods and *relative* probabilities. You can assign a high likelihood to the evidence and still lose probability mass to some other hypothesis, if that other hypothesis assigns a likelihood that is even higher.

Warren seems to be arguing that, given that we see no sabotage, this *confirms* that a Fifth Column exists. You could argue that a Fifth Column *might* delay its sabotage. But the likelihood

1. *http://lesswrong.com/lw/ih/absence_of_evidence_is_evidence_of_absence/*
2. *http://yudkowsky.net/rational/bayes*

is still higher that the *absence* of a Fifth Column would perform an absence of sabotage.

Let E stand for the observation of sabotage, H1 for the hypothesis of a Japanese-American Fifth Column, and H2 for the hypothesis that no Fifth Column exists. Whatever the likelihood that a Fifth Column would do no sabotage, the probability P(E|H1), it cannot be as large as the likelihood that no Fifth Column does no sabotage, the probability P(E|H2). So observing a lack of sabotage increases the probability that no Fifth Column exists.

A lack of sabotage doesn't *prove* that no Fifth Column exists. Absence of *proof* is not *proof* of absence. In logic, A->B, "A implies B", is not equivalent to ~A->~B, "not-A implies not-B".

But in probability theory, absence of *evidence* is always *evidence* of absence. If E is a binary event and P(H|E) > P(H), "seeing E increases the probability of H"; then P(H|~E) < P(H), "failure to observe E decreases the probability of H". P(H) is a weighted mix of P(H|E) and P(H|~E), and necessarily lies between the two. If any of this sounds at all confusing, see An Intuitive Explanation of Bayesian Reasoning[3].

Under the vast majority of real-life circumstances, a cause may not reliably produce signs of itself, but the absence of the cause is even less likely to produce the signs. The absence of an observation may be strong evidence of absence or very weak evidence of absence, depending on how likely the cause is to produce the observation. The absence of an observation that is only weakly permitted (even if the alternative hypothesis does not allow it at all), is very weak evidence of absence (though it is evidence nonetheless). This is the fallacy of "gaps in the fossil record"—fossils form only rarely; it is futile to trumpet the absence of a weakly permitted observation when many strong positive observations have already been recorded. But if there

3. *http://yudkowsky.net/rational/bayes*

are *no* positive observations at all, it is time to worry; hence the Fermi Paradox.

Your strength as a rationalist[4] is your ability to be more confused by fiction than by reality; if you are equally good at explaining any outcome you have zero knowledge. The strength of a model is not what it *can* explain, but what it *can't*, for only prohibitions constrain anticipation[5]. If you don't notice when your model makes the evidence unlikely, you might as well have no model, and also you might as well have no evidence; no brain and no eyes.

4. Page 62, 'Your Strength as a Rationalist'.
5. Page 43, 'Belief in Belief'.

10. Conservation of Expected Evidence[1]

Followup to: Absence of Evidence *Is* Evidence of Absence.[2]

Friedrich Spee von Langenfeld, a priest who heard the confessions of condemned witches, wrote in 1631 the *Cautio Criminalis* ('prudence in criminal cases') in which he bitingly described the decision tree for condemning accused witches: If the witch had led an evil and improper life, she was guilty; if she had led a good and proper life, this too was a proof, for witches dissemble and try to appear especially virtuous. After the woman was put in prison: if she was afraid, this proved her guilt; if she was not afraid, this proved her guilt, for witches characteristically pretend innocence and wear a bold front. Or on hearing of a denunciation of witchcraft against her, she might seek flight or remain; if she ran, that proved her guilt; if she remained, the devil had detained her so she could not get away.

Spee acted as confessor to many witches; he was thus in a position to observe *every* branch of the accusation tree, that no matter *what* the accused witch said or did, it was held a proof against her. In any individual case, you would only hear one branch of the dilemma. It is for this reason that scientists write down their experimental predictions in advance.

But *you can't have it both ways*—as a matter of probability theory, not mere fairness. The rule that "absence of evidence *is* evidence of absence[3]" is a special case of a more general law, which I would name Conservation of Expected Evidence: The *expectation* of the posterior probability, after viewing the evidence, must equal the prior probability.

P(H) = P(H)
P(H) = P(H,E) + P(H,~E)
P(H) = P(H|E)*P(E) + P(H|~E)*P(~E)

1. *http://lesswrong.com/lw/ii/conservation_of_expected_evidence/*
2. Page 65, 'Absence of Evidence Is Evidence of Absence'.
3. Page 65, 'Absence of Evidence Is Evidence of Absence'.

Therefore, for every expectation of evidence, there is an equal and opposite expectation of counterevidence.

If you expect a strong probability of seeing weak evidence in one direction, it must be balanced by a weak expectation of seeing strong evidence in the other direction. If you're very confident in your theory, and therefore anticipate seeing an outcome that matches your hypothesis, this can only provide a very small increment to your belief (it is already close to 1); but the unexpected failure of your prediction would (and must) deal your confidence a huge blow. On *average,* you must expect to be *exactly* as confident as when you started out. Equivalently, the mere *expectation* of encountering evidence—before you've actually seen it—should not shift your prior beliefs. (Again, if this is not intuitively obvious, see An Intuitive Explanation of Bayesian Reasoning[4].)

So if you claim[5] that "no sabotage" is evidence *for* the existence of a Japanese-American Fifth Column, you must conversely hold that seeing sabotage would argue *against* a Fifth Column. If you claim that "a good and proper life" is evidence that a woman is a witch, then an evil and improper life must be evidence that she is not a witch. If you argue[6] that God, to test humanity's faith, refuses to reveal His existence, then the miracles described in the Bible must argue against the existence of God.

Doesn't quite sound right, does it? Pay attention to that feeling of *this seems a little forced,* that quiet strain in the back of your mind[7]. It's important.

For a true Bayesian, it is impossible to seek evidence that *confirms* a theory. There is no possible plan you can devise, no clever strategy, no cunning device, by which you can legitimately expect your confidence in a fixed proposition to be higher (on

4. *http://yudkowsky.net/rational/bayes*
5. Page 65, 'Absence of Evidence Is Evidence of Absence'.
6. *http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/*
7. Page 62, 'Your Strength as a Rationalist'.

average) than before. You can only ever seek evidence to *test* a theory, not to confirm it.

This realization can take quite a load off your mind. You need not worry about how to interpret every possible experimental result to confirm your theory. You needn't bother planning how to make *any* given iota of evidence confirm your theory, because you know that for every expectation of evidence, there is an equal and oppositive expectation of counterevidence. If you try to weaken the counterevidence of a possible "abnormal" observation, you can only do it by weakening the support of a "normal" observation, to a precisely equal and opposite degree. It is a zero-sum game. No matter how you connive, no matter how you argue, no matter how you strategize, you can't possibly expect the resulting game plan to shift your beliefs (on average) in a particular direction.

You might as well sit back and relax while you wait for the evidence to come in.

...human psychology is *so* screwed up.

11. Hindsight bias[1]

Hindsight bias is when people who know the answer vastly overestimate its *predictability* or *obviousness,* compared to the estimates of subjects who must guess without advance knowledge. Hindsight bias is sometimes called the *I-knew-it-all-along effect*.

Fischhoff and Beyth (1975) presented students with historical accounts of unfamiliar incidents, such as a conflict between the Gurkhas and the British in 1814. Given the account as background knowledge, five groups of students were asked what they would have predicted as the probability for each of four outcomes: British victory, Gurkha victory, stalemate with a peace settlement, or stalemate with no peace settlement. Four experimental groups were respectively told that these four outcomes were the historical outcome. The fifth, control group was not told any historical outcome. In every case, a group told an outcome assigned substantially higher probability to that outcome, than did any other group or the control group.

Hindsight bias matters in legal cases, where a judge or jury must determine whether a defendant was legally negligent in failing to foresee a hazard (Sanchiro 2003). In an experiment based on an actual legal case, Kamin and Rachlinski (1995) asked two groups to estimate the probability of flood damage caused by blockage of a city-owned drawbridge. The control group was told only the background information known to the city when it decided not to hire a bridge watcher. The experimental group was given this information, plus the fact that a flood had actually occurred. Instructions stated the city was negligent if the foreseeable probability of flooding was greater than 10%. 76% of the control group concluded the flood was so unlikely that no precautions were necessary; 57% of the experimental group concluded the flood was so likely that failure to take precautions was legally negligent. A third experimental

1. *http://lesswrong.com/lw/il/hindsight_bias/*

group was told the outcome andalso explicitly instructed to avoid hindsight bias, which made no difference: 56% concluded the city was legally negligent.

Viewing history through the lens of hindsight, we vastly underestimate the cost of effective safety precautions. In 1986, the *Challenger* exploded for reasons traced to an O-ring losing flexibility at low temperature. There were warning signs of a problem with the O-rings. But preventing the *Challenger* disaster would have required, not attending to the problem with the O-rings, but attending to *every* warning sign which seemed as severe as the O-ring problem, *without benefit of hindsight*. It could have been done, but it would have required a *general policy* much more expensive than just fixing the O-Rings.

Shortly after September 11th 2001, I thought to myself, *and now someone will turn up minor intelligence warnings of something-or-other, and then the hindsight will begin*. Yes, I'm sure they had some minor warnings of an al Qaeda plot, but they probably also had minor warnings of mafia activity, nuclear material for sale, and an invasion from Mars.

Because we don't see the cost of a general policy, we learn overly specific lessons. After September 11th, the FAA prohibited box-cutters on airplanes—as if the problem had been the failure to take *this particular* "obvious" precaution. We don't learn the general lesson: *the cost of effective caution is very high because you must attend to problems that are not as obvious now as past problems seem in hindsight*.

The test of a model is how much probability it assigns to the observed outcome. Hindsight bias systematically distorts this test; we think our model assigned much more probability than it actually did. Instructing the jury doesn't help. You have to write down your predictions in advance[2]. Or as Fischhoff (1982) put it:

2. Page 68, 'Conservation of Expected Evidence'.

> When we attempt to understand past events, we implicitly test the hypotheses or rules we use both to interpret and to anticipate the world around us. If, in hindsight, we systematically underestimate the surprises that the past held and holds for us, we are subjecting those hypotheses to inordinately weak tests and, presumably, finding little reason to change them.

Fischhoff, B. 1982. For those condemned to study the past: Heuristics and biases in hindsight. In Kahneman et. al. 1982: 332–351.

Fischhoff, B., and Beyth, R. 1975. I knew it would happen: Remembered probabilities of once-future things. Organizational Behavior and Human Performance, 13: 1-16.

Kamin, K. and Rachlinski, J. 1995. Ex Post ≠ Ex Ante: Determining Liability in Hindsight[3]. Law and Human Behavior, 19(1): 89-104.

Sanchiro, C. 2003. Finding Error. Mich. St. L. Rev. 1189.

3. *http://www.jstor.org/view/01477307/ap050075/05a00120/0*

12. Hindsight Devalues Science[1]

This excerpt[2] from Meyers's *Exploring Social Psychology* is worth reading in entirety. Cullen Murphy, editor of *The Atlantic,* said that the social sciences turn up "no ideas or conclusions that can't be found in [any] encyclopedia of quotations... Day after day social scientists go out into the world. Day after day they discover that people's behavior is pretty much what you'd expect."

Of course, the "expectation" is all hindsight[3]. (Hindsight bias: Subjects who know the actual answer to a question assign much higher probabilities they "would have" guessed for that answer, compared to subjects who must guess without knowing the answer.)

The historian Arthur Schlesinger, Jr. dismissed scientific studies of WWII soldiers' experiences as "ponderous demonstrations" of common sense. For example:

1. Better educated soldiers suffered more adjustment problems than less educated soldiers. (Intellectuals were less prepared for battle stresses than street-smart people.)
2. Southern soldiers coped better with the hot South Sea Island climate than Northern soldiers. (Southerners are more accustomed to hot weather.)
3. White privates were more eager to be promoted to noncommissioned officers than Black privates. (Years of oppression take a toll on achievement motivation.)
4. Southern Blacks preferred Southern to Northern White officers (because Southern officers were more experienced and skilled in interacting with Blacks).
5. As long as the fighting continued, soldiers were more eager to return home than after the war ended.

1. *http://lesswrong.com/lw/im/hindsight_devalues_science/*
2. *http://csml.som.ohio-state.edu/Music829C/hindsight.bias.html*
3. Page 71, 'Hindsight bias'.

(During the fighting, soldiers knew they were in mortal danger.)

How many of these findings do you think you *could have* predicted in advance? 3 out of 5? 4 out of 5? Are there any cases where you would have predicted the opposite—where your model takes a hit[4]? Take a moment to think before continuing...

In this demonstration (from Paul Lazarsfeld by way of Meyers), all of the findings above are the *opposite* of what was actually found. How many times did you think your model took a hit? How many times did you admit you would have been wrong? That's how good your model really was. The measure of your strength as a rationalist[5] is your ability to be more confused by fiction than by reality.

Unless, of course, I reversed the results again. What do you think?

Do your thought processes at this point, where you *really don't* know the answer, feel different from the thought processes you used to rationalize either side of the "known" answer?

Daphna Baratz exposed college students to pairs of supposed findings, one true ("In prosperous times people spend a larger portion of their income than during a recession") and one the truth's opposite. In both sides of the pair, students rated the supposed finding as what they "would have predicted". Perfectly standard hindsight bias.

Which leads people to think they have no need for science, because they "could have predicted" that.

(Just as you would expect, right?)

Hindsight will lead us to systematically undervalue the surprisingness of scientific findings, especially the discoveries we *understand*—the ones that seem real to us, the ones we can retrofit into our models of the world. If you understand neurology or physics and read news in that topic, then you probably

4. Page 68, 'Conservation of Expected Evidence'.
5. Page 62, 'Your Strength as a Rationalist'.

underestimate the surprisingness of findings in those fields too. This unfairly devalues the contribution of the researchers; and worse, will prevent you from noticing when you are seeing evidence that doesn't fit[6] what you *really* would have expected.

We need to make a conscious effort to be shocked *enough*.

6. Page 68, 'Conservation of Expected Evidence'.

13. Fake Explanations[1]

Once upon a time, there was an instructor who taught physics students. One day she called them into her class, and showed them a wide, square plate of metal, next to a hot radiator. The students each put their hand on the plate, and found the side next to the radiator cool, and the distant side warm. And the instructor said, *Why do you think this happens?* Some students guessed convection of air currents, and others guessed strange metals in the plate. They devised many creative explanations, none stooping so low as to say "I don't know" or "This seems impossible.[2]"

And the answer was that before the students entered the room, the instructor turned the plate around.

Consider the student who frantically stammers, "Eh, maybe because of the heat conduction and so?" I ask: is this answer a proper belief[3]? The words are easily enough professed[4]—said in a loud, emphatic voice. But do the words actually control anticipation[5]?

Ponder that innocent little phrase, "because of", which comes before "heat conduction". Ponder some of the *other* things we could put after it. We could say, for example, "Because of phlogiston", or "Because of magic."

"Magic!" you cry. "That's not a *scientific* explanation!" Indeed, the phrases "because of heat conduction" and "because of magic" are readily recognized as belonging to different *literary genres*. "Heat conduction" is something that Spock might say on *Star Trek*, whereas "magic" would be said by Giles in *Buffy the Vampire Slayer*.

---

1. *http://lesswrong.com/lw/ip/fake_explanations/*
2. Page 62, 'Your Strength as a Rationalist'.
3. Page 53, 'Belief as Attire'.
4. Page 50, 'Professing and Cheering'.
5. Page 43, 'Belief in Belief'.

However, as Bayesians, we take no notice of literary genres. For us, the substance of a model is the control it exerts on anticipation. If you say "heat conduction", what experience does that lead you to *anticipate?* Under normal circumstances, it leads you to anticipate that, if you put your hand on the side of the plate near the radiator, that side will feel warmer than the opposite side. If "because of heat conduction" can also explain the radiator-adjacent side feeling *cooler,* then it can explain pretty much *anything*.

And[6] as[7] we[8] all[9] know[10] by[11] this[12] point[13] (I[14] do[15] hope[16]), if you are equally good at explaining any outcome, you have zero knowledge. "Because of heat conduction", used in such fashion, is a disguised hypothesis of maximum entropy. It is anticipation-isomorphic to saying "magic". It feels like an explanation, but it's not.

Supposed that instead of guessing, we measured the heat of the metal plate at various points and various times. Seeing a metal plate next to the radiator, we would ordinarily expect the point temperatures to satisfy an equilibrium of the diffusion equation with respect to the boundary conditions imposed by the environment. You might not know the exact temperature of the first point measured, but after measuring the first points—I'm not physicist enough to know how many would be required—you could take an excellent guess at the rest.

A true master of the art of using numbers to constrain the anticipation of material phenomena—a "physicist"—would take some measurements and say, "This plate was in equilibrium

6. Page 39, 'Making Beliefs Pay Rent (in Anticipated Experiences)'.
7. Page 43, 'Belief in Belief'.
8. Page 48, 'Bayesian Judo'.
9. Page 50, 'Professing and Cheering'.
10. Page 53, 'Belief as Attire'.
11. *http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/*
12. Page 55, 'Focus Your Uncertainty'.
13. Page 62, 'Your Strength as a Rationalist'.
14. Page 65, 'Absence of Evidence Is Evidence of Absence'.
15. Page 74, 'Hindsight Devalues Science'.
16. Page 68, 'Conservation of Expected Evidence'.

with the environment two and a half minutes ago, turned around, and is now approaching equilibrium again."

The deeper error of the students is not simply that they failed to constrain anticipation. Their deeper error is that they thought they were doing physics. They said the phrase "because of", followed by the sort of words Spock might say on *Star Trek,* and thought they thereby entered the magisterium of science.

Not so. They simply moved their magic from one literary genre to another.

14. Guessing the Teacher's Password[1]

Followup to: Fake Explanations[2]

When I was young, I read popular physics books such as Richard Feynman's *QED: The Strange*[3] *Theory of Light and Matter*. I knew that light was waves, sound was waves, matter was waves. I took pride in my scientific literacy, when I was nine years old.

When I was older, and I began to read the *Feynman Lectures on Physics,* I ran across a gem called "the wave equation". I could follow the equation's derivation, but, looking back[4], I couldn't see its truth at a glance. So I thought about the wave equation for three days, on and off, until I saw that it was embarrassingly obvious. And when I finally understood, I realized that the whole time I had accepted the honest assurance of physicists that light was waves, sound was waves, matter was waves, I had not had the vaguest idea of what the word "wave" meant to a physicist.

There is an instinctive tendency to think that if a physicist says "light is made of waves", and the teacher says "What is light made of?", and the student says "Waves!", the student has made a true statement. That's only fair, right? We accept "waves" as a correct answer from the physicist; wouldn't it be unfair to reject it from the student? Surely, the answer "Waves!" is either *true* or *false,* right?

Which is one more bad habit to unlearn from school[5]. Words do not have intrinsic definitions. If I hear the syllables "bea-ver" and think of a large rodent, that is a fact about my own state of mind, not a fact about the syllables "bea-ver". The sequence of syllables "made of waves" (or "because of heat

1. *http://lesswrong.com/lw/iq/guessing_the_teachers_password/*
2. Page 77, 'Fake Explanations'.
3. *http://lesswrong.com/lw/hs/think_like_reality/*
4. *http://www.math.utah.edu/~pa/math/polya.html*
5. *http://lesswrong.com/lw/i2/two_more_things_to_unlearn_from_school/*

conduction[6]") is not a *hypothesis,* it is a pattern of vibrations traveling through the air, or ink on paper. It can *associate* to a hypothesis in someone's mind, but it is not, of itself, right or wrong. But in school, the teacher hands you a gold star for *saying* "made of waves", which must be the correct answer because the teacher heard a physicist emit the same sound-vibrations. Since verbal behavior (spoken or written) is what gets the gold star, students begin to think that verbal behavior has a truth-value. After all, either light is made of waves, or it isn't, right?

And this leads into an even worse habit. Suppose the teacher presents you with a confusing problem[7] involving a metal plate next to a radiator; the far side feels warmer than the side next to the radiator. The teacher asks "Why?" If you say "I don't know", you have *no* chance of getting a gold star—it won't even count as class participation. But, during the current semester, this teacher has used the phrases "because of heat convection", "because of heat conduction", and "because of radiant heat". One of these is probably what the teacher wants. You say, "Eh, maybe because of heat conduction?"

This is not a hypothesis *about* the metal plate. This is not even a proper belief[8]. It is an attempt to *guess the teacher's password.*

Even visualizing the symbols of the diffusion equation (the math governing heat conduction) doesn't mean you've formed a hypothesis *about* the metal plate. This is not school; we are not testing your memory to see if you can write down the diffusion equation. This is Bayescraft; we are scoring your anticipations of experience. If you *use* the diffusion equation, by measuring a few points with a thermometer and then trying to predict what the thermometer will say on the next measurement, then it is definitely connected to experience. Even if the student just visualizes something *flowing,* and therefore holds a match near

6. Page 77, 'Fake Explanations'.
7. Page 77, 'Fake Explanations'.
8. Page 53, 'Belief as Attire'.

the cooler side of the plate to try to measure where the heat goes, then this mental image of flowing-ness connects to experience; it controls anticipation.

If you aren't *using* the diffusion equation—putting in numbers and getting out results that control your anticipation of particular experiences—then the connection between map and territory is severed as though by a knife. What remains is not a belief[9], but a verbal behavior.

In the school system, it's all about verbal behavior, whether written on paper or spoken aloud. Verbal behavior gets you a gold star or a failing grade. Part of unlearning this bad habit is becoming consciously aware of the difference between an explanation and a password.

Does this seem too harsh? When you're faced by a confusing metal plate, can't "Heat conduction?" be a first step toward finding the answer? Maybe, but only if you don't fall into the trap of thinking that you are looking for a password. What if there is no teacher to tell you that you failed? Then you may think that "Light is wakalixes" is a good explanation, that "wakalixes" is the correct password. It happened to me when I was nine years old—not because I was stupid, but because this is what happens *by default*. This is how human beings think, unless they are trained *not* to fall into the trap. Humanity stayed stuck in holes like this for thousands of years.

Maybe, if we drill students that *words don't count, only anticipation-controllers,* the student will *not* get stuck on "Heat conduction? No? Maybe heat convection? That's not it either?" Maybe *then,* thinking the phrase "Heat conduction" will lead onto a genuinely helpful path, like:

- "Heat conduction?"
- But that's only a phrase—what does it mean?
- The diffusion equation?
- But those are only symbols—how do I apply them?

9. Page 53, 'Belief as Attire'.

- What does applying the diffusion equation lead me to anticipate?
- It sure doesn't lead me to anticipate that the side of a metal plate farther away from a radiator would feel warmer.
- I notice[10] that I am confused[11]. Maybe the near side just *feels* cooler, because it's made of more insulative material and transfers less heat to my hand? I'll try measuring the temperature...
- Okay, that wasn't it. Can I try to verify whether the diffusion equation holds true of this metal plate, at all? Is heat *flowing* the way it usually does, or is something else going on?
- I could hold a match to the plate and try to measure how heat spreads over time...

If we are *not* strict about "Eh, maybe because of heat conduction?" being a fake explanation, the student will very probably get stuck on some wakalixes-password. *This happens by default, it happened to the whole human species for thousands of years.*

10. Page 62, 'Your Strength as a Rationalist'.
11. Page 74, 'Hindsight Devalues Science'.

15. Science as Attire[1]

Prerequisites: Fake Explanations[2], Belief As Attire[3]

The preview for the *X-Men* movie has a voice-over saying: "In every human being... there is the genetic code... for mutation." Apparently you can acquire all sorts of neat abilities by mutation. The mutant Storm, for example, has the ability to throw lightning bolts.

I beg you, dear reader, to consider the biological machinery necessary to generate electricity; the biological adaptations necessary to avoid being harmed by electricity; and the cognitive circuitry required for finely tuned control of lightning bolts. If we actually observed any organism acquiring these abilities *in one generation,* as the result of *mutation,* it would outright falsify the neo-Darwinian model of natural selection. It would be worse than finding rabbit fossils in the pre-Cambrian. If evolutionary theory could *actually* stretch to cover Storm, it would be able to explain anything[4], and we all know what that would imply.

The *X-Men* comics use terms like "evolution", "mutation", and "genetic code", purely to place themselves in what they conceive to be the *literary genre* of science. The part that scares me is wondering how many people, especially in the media, understand science *only* as a literary genre.

---

1. *http://lesswrong.com/lw/ir/science_as_attire/*
2. Page 77, 'Fake Explanations'.
3. Page 53, 'Belief as Attire'.
4. Page 62, 'Your Strength as a Rationalist'.

I encounter people who very definitely believe in[5] evolution, who sneer at the folly of creationists. And yet they have no idea of what the theory of evolutionary biology permits and prohibits. They'll talk about "the next step in the evolution of humanity", as if natural selection got here by following a plan. Or even worse, they'll talk about something completely outside the domain of evolutionary biology, like an improved design for computer chips, or corporations splitting, or humans uploading themselves into computers, and they'll call *that* "evolution". If evolutionary biology could cover that, it could cover anything.

Probably an actual majority of the people who *believe in* evolution use the phrase "because of evolution[6]" because they want to be part of the scientific in-crowd—belief as scientific attire[7], like wearing a lab coat. If the scientific in-crowd instead used the phrase "because of intelligent design", they would just as cheerfully use that instead—it would make no difference to their anticipation-controllers. Saying "because of evolution" instead of "because of intelligent design" does not, *for them,* prohibit Storm. Its only purpose, for them, is to identify with a tribe.

I encounter people who are quite willing to entertain the notion of dumber-than-human Artificial Intelligence, or even mildly smarter-than-human Artificial Intelligence. Introduce the notion of strongly superhuman Artificial Intelligence, and they'll suddenly decide it's "pseudoscience[8]". It's not that they think they have a theory of intelligence which lets them calculate a theoretical upper bound on the power of an optimization process. Rather, they associate strongly superhuman AI to the *literary genre* of apocalyptic literature; whereas an AI running a small corporation associates to the literary genre of *Wired* magazine. They aren't speaking from within a model of cog-

5. Page 50, 'Professing and Cheering'.
6. Page 77, 'Fake Explanations'.
7. Page 53, 'Belief as Attire'.
8. *http://lesswrong.com/lw/io/is_molecular_nanotechnology_scientific/*

nition. They don't realize they *need* a model. They don't realize that science is *about* models. Their devastating critiques consist purely of *comparisons to apocalyptic literature*, rather than, say, known laws which prohibit such an outcome. They understand science *only* as a literary genre, or in-group to belong to. The attire[9] doesn't look to them like a lab coat; this isn't the football team they're cheering[10] for.

Is there anything in science that you are *proud* of believing, and yet you do not use the belief professionally? You had best ask yourself which future experiences your belief *prohibits* from happening to you. That is the sum of what you have assimilated and made a true part of yourself. Anything else is probably passwords[11] or attire[12].

9. Page 53, 'Belief as Attire'.
10. Page 50, 'Professing and Cheering'.
11. Page 80, 'Guessing the Teacher's Password'.
12. Page 53, 'Belief as Attire'.

16. Fake Causality[1]

Followup to: Fake Explanations[2], Guessing the Teacher's Password[3]

Phlogiston was the 18 century's answer to the Elemental Fire of the Greek alchemists. Ignite wood, and let it burn. What is the orangey-bright "fire" stuff? Why does the wood transform into ash? To both questions, the 18th-century chemists answered, "phlogiston".

...and that was it, you see, that was their answer: "Phlogiston."

Phlogiston escaped from burning substances as visible fire. As the phlogiston escaped, the burning substances lost phlogiston and so became ash, the "true material". Flames in enclosed containers went out because the air became saturated with phlogiston, and so could not hold any more. Charcoal left little residue upon burning because it was nearly pure phlogiston.

Of course, one didn't use phlogiston theory to *predict* the outcome of a chemical transformation. You looked at the result first, then you used phlogiston theory to *explain* it. It's not that phlogiston theorists predicted a flame would extinguish in a closed container; rather they lit a flame in a container, watched it go out, and then said, "The air must have become saturated with phlogiston." You couldn't even use phlogiston theory to say what you ought *not* to see[4]; it could explain everything.

This was an earlier age of science. For a long time, no one realized there was a problem. Fake explanations[5] don't *feel* fake. That's what makes them dangerous.

---

1. *http://lesswrong.com/lw/is/fake_causality/*
2. Page 77, 'Fake Explanations'.
3. Page 80, 'Guessing the Teacher's Password'.
4. Page 62, 'Your Strength as a Rationalist'.
5. Page 77, 'Fake Explanations'.

Modern research suggests that humans think about cause and effect using something like the directed acyclic graphs (DAGs) of Bayes nets. Because it rained, the sidewalk is wet; because the sidewalk is wet, it is slippery:

[Rain] -> [Sidewalk wet] -> [Sidewalk slippery]

From this we can infer—or, in a Bayes net, rigorously calculate in probabilities—that when the sidewalk is slippery, it probably rained; but if we already know that the sidewalk is wet, learning that the sidewalk is slippery tells us nothing more about whether it rained.

Why is fire hot and bright when it burns?

["Phlogiston"] -> [Fire hot and bright]

It *feels* like an explanation. It's *represented* using the same cognitive data format. But the human mind does not automatically detect when a cause has an unconstraining arrow to its effect. Worse, thanks to hindsight bias[6], it may feel like the cause constrains[7] the effect, when it was merely fitted[8] to the effect.

Interestingly, our modern understanding of probabilistic reasoning about causality[9] can describe precisely what the phlogiston theorists were doing wrong. One of the primary inspirations for Bayesian networks was noticing the problem of double-counting evidence if inference resonates between an effect and a cause. For example, let's say that I get a bit of unreliable information that the sidewalk is wet. This should make me think it's more likely to be raining. But, if it's more likely to be raining, doesn't that make it more likely that the sidewalk is wet? And wouldn't *that* make it more likely that the sidewalk is

6. Page 71, 'Hindsight bias'.
7. Page 39, 'Making Beliefs Pay Rent (in Anticipated Experiences)'.
8. Page 68, 'Conservation of Expected Evidence'.
9. *http://books.google.com/books?id=k9VsqN24pNYC&dq=&pg=PP1&ots=WR9UGWdOdd&sig=w_Mrax-y4VVwZy5SQGySphNsKMc&prev=http://www.google.com/search%3Fhl%3Den%26safe%3Doff%26q%3Dpearl%2Bintelligent%2Bsystems%26btnG%3DSearch&s*

slippery? But if the sidewalk is slippery, it's probably wet; and then I should again raise my probability that it's raining...

Judea Pearl uses the metaphor of an algorithm for counting soldiers in a line. Suppose you're in the line, and you see two soldiers next to you, one in front and one in back. That's three soldiers. So you ask the soldier next to you, "How many soldiers do *you* see?" He looks around and says, "Three". So that's a total of six soldiers. This, obviously, is *not* how to do it.

A smarter way is to ask the soldier in front of you, "How many soldiers forward of you?" and the soldier in back, "How many soldiers backward of you?" The question "How many soldiers forward?" can be passed on as a message without confusion. If I'm at the front of the line, I pass the message "1 soldier forward", for myself. The person directly in back of me gets the message "1 soldier forward", and passes on the message "2 soldiers forward" to the soldier behind him. At the same time, each soldier is also getting the message "N soldiers backward" from the soldier behind them, and passing it on as "N+1 soldiers backward" to the soldier in front of them. How many soldiers in total? Add the two numbers you receive, plus one for yourself: that is the total number of soldiers in line.

The key idea is that every soldier must *separately* track the two messages, the forward-message and backward-message, and add them together only at the end. You never add any soldiers from the backward-message you receive to the forward-message you pass back. Indeed, the total number of soldiers is never passed as a message—no one ever says it aloud.

An analogous principle operates in rigorous probabilistic reasoning about causality. If you learn something about whether it's raining, from some source *other* than observing the sidewalk to be wet, this will send a forward-message from [rain] to [sidewalk wet] and raise our expectation of the sidewalk being wet. If you observe the sidewalk to be wet, this sends a backward-message to our belief that it is raining, and this message propagates from [rain] to all neighboring nodes *except* the

[sidewalk wet] node. We count each piece of evidence exactly once; no update message ever "bounces" back and forth. The exact algorithm may be found in Judea Pearl's classic "Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference[10]".

So what went wrong in phlogiston theory? When we observe that fire is hot, the [fire] node can send a backward-evidence to the ["phlogiston"] node, leading us to update our beliefs about phlogiston. But if so, we can't count this as a successful forward-prediction of phlogiston theory. The message should go in only one direction, and not bounce back.

Alas, human beings do not use a rigorous algorithm for updating belief networks. We learn about parent nodes from observing children, and predict child nodes from beliefs about parents. But we don't keep rigorously separate books for the backward-message and forward-message. We just remember that phlogiston is hot, which *causes* fire to be hot. So it seems like phlogiston theory predicts the hotness of fire. Or, worse, it just feels like *phlogiston makes the fire hot*.

Until you notice that no *advance* predictions are being made, the non-constraining causal node is not labeled "fake". It's represented the same way as any other node in your belief network. It feels like a fact, like all the other facts you know: *Phlogiston makes the fire hot*.

A properly designed AI would notice the problem instantly. This wouldn't even require special-purpose code, just correct bookkeeping of the belief network. (Sadly, we humans can't rewrite our own code, the way a properly designed AI could.)

Speaking of "hindsight bias[11]" is just the nontechnical way of saying that humans do not rigorously separate forward and

10. *http://books.google.com/books?id=k9VsqN24pNYC&dq=&pg=PP1&ots=WR9UGWdOdd&sig=w_Mrax-y4VVwZy5SQGySphNsKMc&prev=http://www.google.com/search%3Fhl%3Den%26safe%3Doff%26q%3Dpearl%2Bintelligent%2Bsystems%26btnG%3DSearch&s*

11. Page 74, 'Hindsight Devalues Science'.

backward messages, allowing forward messages to be contaminated by backward ones.

Those who long ago went down the path of phlogiston were not trying to be fools. No scientist deliberately wants to get stuck in a blind alley. Are there any fake explanations in *your* mind? If there are, I guarantee they're not labeled "fake explanation", so polling your thoughts for the "fake" keyword will not turn them up.

Thanks to hindsight bias[12], it's also not enough to check how well your theory "predicts" facts you already know. You've got to predict for tomorrow, not yesterday. It's the only way a messy human mind can be guaranteed of sending a pure forward message.

12. Page 74, 'Hindsight Devalues Science'.

17. Semantic Stopsigns[1]

And the child asked:

Q: Where did this rock come from?
A: I chipped it off the big boulder, at the center of the village.
Q: Where did the boulder come from?
A: It probably rolled off the huge mountain that towers over our village.
Q: Where did the mountain come from?
A: The same place as all stone: it is the bones of Ymir, the primordial giant.
Q: Where did the primordial giant, Ymir, come from?
A: From the great abyss, Ginnungagap.
Q: Where did the great abyss, Ginnungagap, come from?
A: Never ask that question.

Consider the seeming paradox of the First Cause. Science has traced events back to the Big Bang, but why did the Big Bang happen? It's all well and good to say that the zero of time begins at the Big Bang—that there is nothing before the Big Bang in the ordinary flow of minutes and hours. But saying this presumes our physical law, which itself appears highly structured; it calls out for explanation. Where did the physical laws come from? You could say that we're all a computer simulation, but then the computer simulation is running on some other world's laws of physics—where did *those* laws of physics come from?

At this point, some people say, "God!"

What could possibly make anyone, even a highly religious person, think this even *helped* answer the paradox of the First Cause? Why wouldn't you automatically ask, "Where did God come from?" Saying "God is uncaused" or "God created Himself" leaves us in exactly the same position as "Time began with the Big Bang." We just ask why the whole metasystem exists in

1. *http://lesswrong.com/lw/it/semantic_stopsigns/*

the first place, or why some events but not others are allowed to be uncaused.

My purpose here is not to discuss the seeming paradox of the First Cause, but to ask why anyone would think "God!" *could* resolve the paradox. Saying "God!" is a way of belonging to a tribe[2], which gives people a motive to say it as often as possible—some people even say it for questions like "Why did this hurricane strike New Orleans?" Even so, you'd hope people would notice that on the *particular* puzzle of the First Cause, saying "God!" doesn't help. It doesn't make the paradox seem any less paradoxical *even if true*. How could anyone *not* notice this?

Jonathan Wallace suggested that "God!" functions as a *semantic stopsign*—that it isn't a propositional assertion, so much as a cognitive traffic signal: do not think past this point. Saying "God!" doesn't so much resolve the paradox, as put up a cognitive traffic signal to halt the obvious continuation of the question-and-answer chain.

Of course *you'd* never do that, being a good and proper atheist, right? But "God!" isn't the *only* semantic stopsign, just the obvious first example.

The transhuman technologies—molecular nanotechnology, advanced biotech, genetech, Artificial Intelligence, et cetera—pose tough policy questions. What kind of role, if any, should a government take in supervising a parent's choice of genes for their child? Could parents deliberately choose genes for schizophrenia? If enhancing a child's intelligence is expensive, should governments help ensure access, to prevent the emergence of a cognitive elite? You can propose various institutions to answer these policy questions—for example, that private charities should provide financial aid for intelligence enhancement—but the obvious next question is, "Will this institution be effective?" If we rely on product liability lawsuits to

2. Page 50, 'Professing and Cheering'.

prevent corporations from building harmful nanotech, will that really *work?*

I know someone whose answer to every one of these questions is "Liberal democracy!" That's it. That's his answer. If you ask the obvious question of "How well have liberal democracies performed, historically, on problems this tricky?" or "What if liberal democracy does something stupid?" then you're an autocrat, or libertopian, or otherwise a very very bad person. No one is allowed to question democracy.

I once called this kind of thinking "the divine right of democracy". But it is more precise to say that "Democracy!" functioned for him as a semantic stopsign. If anyone had said to him "Turn it over to the Coca-Cola corporation!", he would have asked the obvious next questions: "Why? What will the Coca-Cola corporation do about it? Why should we trust them? Have they done well in the past on equally tricky problems?"

Or suppose that someone says "Mexican-Americans are plotting to remove all the oxygen in Earth's atmosphere." You'd probably ask, "Why would they do *that?* Don't Mexican-Americans have to breathe too? Do Mexican-Americans even function as a unified conspiracy?" If you don't ask these obvious next questions when someone says, "Corporations are plotting to remove Earth's oxygen," then "Corporations!" functions for you as a semantic stopsign.

Be careful here not to create a new generic counterargument against things you don't like—"Oh, it's just a stopsign!" No word is a stopsign of itself; the question is whether a word has that effect on a particular person. Having strong emotions[3] about something doesn't qualify it as a stopsign. I'm not exactly fond of terrorists or fearful of private property; that doesn't mean "Terrorists!" or "Capitalism!" are cognitive traffic signals unto me. (The word "intelligence" did once have that effect on

3. Page 463, 'Feeling Rational'.

me, though no longer.) What distinguishes a semantic stopsign is *failure to consider the obvious next question*.

18. Mysterious Answers to Mysterious Questions[1]

Imagine looking at your hand, and knowing nothing of cells, nothing of biochemistry, nothing of DNA. You've learned some anatomy from dissection, so you know your hand contains muscles; but you don't know why muscles move instead of lying there like clay. Your hand is just... stuff... and for some reason it moves under your direction. Is this not magic?

> "The animal body does not act as a thermodynamic engine ... consciousness teaches every individual that they are, to some extent, subject to the direction of his will. It appears therefore that animated creatures have the power of immediately applying to certain moving particles of matter within their bodies, forces by which the motions of these particles are directed to produce derived mechanical effects... The influence of animal or vegetable life on matter is infinitely beyond the range of any scientific inquiry hitherto entered on. Its power of directing the motions of moving particles, in the demonstrated daily miracle of our human free-will, and in the growth of generation after generation of plants from a single seed, are infinitely different from any possible result of the fortuitous concurrence of atoms... Modern biologists were coming once more to the acceptance of something and that was a vital principle."
>
> -- Lord Kelvin

This was the theory of *vitalism*; that the mysterious difference between living matter and non-living matter was explained by an *elan vital* or *vis vitalis*. Elan vital infused living matter and caused it to move as consciously directed. Elan vital

1. *http://lesswrong.com/lw/iu/mysterious_answers_to_mysterious_questions/*

participated in chemical transformations which no mere non-living particles could undergo—Wöhler's later synthesis of urea, a component of urine, was a major blow to the vitalistic theory because it showed that mere *chemistry* could duplicate a product of biology.

Calling "elan vital" an explanation, even a fake explanation[2] like phlogiston[3], is probably giving it too much credit. It functioned primarily as a curiosity-stopper[4]. You said "Why?" and the answer was "Elan vital!"

When you say "Elan vital!", it *feels* like you know why your hand moves. You have a little causal diagram[5] in your head that says ["Elan vital!"] -> [hand moves]. But actually you know nothing you didn't know before. You don't know, say, whether your hand will generate heat or absorb heat, unless you have observed the fact already; if not, you won't be able to predict it in advance. Your curiosity feels sated, but it hasn't been fed. Since you can say "Why? Elan vital!" to any possible observation, it is equally good at explaining all outcomes, a disguised hypothesis of maximum entropy, etcetera.

But the greater lesson lies in the vitalists' reverence for the elan vital, their eagerness to pronounce it a mystery beyond all science. Meeting the great dragon Unknown, the vitalists did not draw their swords to do battle, but bowed their necks in submission. They took pride[6] in their ignorance, made biology into a *sacred* mystery, and thereby became loath to relinquish their ignorance[7] when evidence came knocking.

The Secret of Life was *infinitely beyond the reach of science!* Not just a *little* beyond, mind you, but *infinitely* beyond! Lord

2. Page 77, 'Fake Explanations'.
3. Page 87, 'Fake Causality'.
4. Page 92, 'Semantic Stopsigns'.
5. Page 87, 'Fake Causality'.
6. *http://lesswrong.com/lw/h8/tsuyoku_naritai_i_want_to_become_stronger/*
7. *http://yudkowsky.net/virtues/*

Kelvin sure did get a tremendous emotional kick out of *not knowing something*.

But ignorance exists in the map, not in the territory. If I am ignorant about a phenomenon, that is a fact about my own state of mind, not a fact about the phenomenon itself. A phenomenon can *seem* mysterious to some particular person. There are no phenomena which are mysterious of themselves. To worship a phenomenon because it seems so wonderfully mysterious, is to worship your own ignorance.

Vitalism shared with phlogiston the error of *encapsulating the mystery as a substance*. Fire was mysterious, and the phlogiston theory encapsulated the mystery in a mysterious substance called "phlogiston". Life was a sacred mystery, and vitalism encapsulated the sacred mystery in a mysterious substance called "elan vital". Neither answer helped concentrate the model's probability density[8]—make some outcomes easier to explain than others. The "explanation" just wrapped up the question as a small, hard, opaque black ball.

In a comedy written by Moliere, a physician explains the power of a soporific by saying that it contains a "dormitive potency". Same principle. It is a failure of human psychology that, faced with a mysterious phenomenon, we more readily postulate mysterious inherent substances than complex underlying processes.

But the deeper failure is supposing that an *answer* can be mysterious. If a phenomenon feels mysterious, that is a fact about our state of knowledge, not a fact about the phenomenon itself. The vitalists saw a mysterious gap in their knowledge, and postulated a mysterious stuff that plugged the gap. In doing so, they mixed up the map with the territory. All confusion and bewilderment exist in the mind, not in encapsulated substances.

This is the ultimate and fully general explanation for why, again and again in humanity's history, people are shocked to

8. Page 55, 'Focus Your Uncertainty'.

discover that an incredibly mysterious question has a non-mysterious answer. Mystery is a property of questions, not answers.

Therefore I call theories such as vitalism *mysterious answers to mysterious questions*.

These are the signs of mysterious answers to mysterious questions:

- First, the explanation acts as a curiosity-stopper[9] rather than an anticipation-controller[10].
- Second, the hypothesis has no moving parts—the model is not a specific complex mechanism, but a blankly solid substance or force. The mysterious substance or mysterious force may be said to be here or there, to cause [11]this or that; but the reason why the mysterious force behaves thus is wrapped in a blank unity.
- Third, those who proffer the explanation cherish their ignorance[12]; they speak proudly of how the phenomenon defeats ordinary science or is unlike merely mundane phenomena.
- Fourth, *even after the answer is given, the phenomenon is still a mystery* and possesses the same quality of wonderful inexplicability that it had at the start.

9. Page 92, 'Semantic Stopsigns'.
10. Page 39, 'Making Beliefs Pay Rent (in Anticipated Experiences)'.
11. Page 87, 'Fake Causality'.
12. *http://yudkowsky.net/virtues/*

19. The Futility of Emergence[1]

Prerequisites: Belief in Belief[2], Fake Explanations[3], Fake Causality[4], Mysterious Answers to Mysterious Questions[5]

The failures of phlogiston[6] and vitalism[7] are historical[8] hindsight[9]. Dare I step out on a limb, and name some *current* theory which I deem analogously flawed?

I name *emergence* or *emergent phenomena*—usually defined as the study of systems whose high-level behaviors arise or "emerge" from the interaction of many low-level elements. (Wikipedia[10]: "The way complex systems and patterns arise out of a multiplicity of relatively simple interactions".) Taken literally, that description fits every phenomenon in our universe above the level of individual quarks, which is part of the problem. Imagine pointing to a market crash and saying "It's not a quark!" Does that feel like an explanation? No? Then neither should saying "It's an emergent phenomenon!"

It's the noun "emergence" that I protest, rather than the verb "emerges from". There's nothing wrong with saying "X emerges from Y", where Y is some specific, detailed model with internal moving parts. "Arises from" is another legitimate phrase that means exactly the same thing: Gravity arises from the curvature of spacetime, according to the specific mathematical model of General Relativity. Chemistry arises from interactions between atoms, according to the specific model of quantum electrodynamics.

---

1. *http://lesswrong.com/lw/iv/the_futility_of_emergence/*
2. Page 43, 'Belief in Belief'.
3. Page 77, 'Fake Explanations'.
4. Page 87, 'Fake Causality'.
5. Page 96, 'Mysterious Answers to Mysterious Questions'.
6. Page 87, 'Fake Causality'.
7. Page 96, 'Mysterious Answers to Mysterious Questions'.
8. Page 74, 'Hindsight Devalues Science'.
9. Page 71, 'Hindsight bias'.
10. *http://en.wikipedia.org/wiki/Emergence*

Now suppose I should say that gravity is explained by "arisence" or that chemistry is an "arising phenomenon", and claim that as my explanation.

The phrase "emerges from" is acceptable, just like "arises from" or "is caused by" are acceptable, if the phrase precedes some specific model to be judged on its own merits.

However, this is *not* the way "emergence" is commonly used. "Emergence" is commonly used as an explanation in its own right.

I have lost track of how many times I have heard people say, "Intelligence is an emergent phenomenon!" as if that explained intelligence. This usage fits all the checklist items for a mysterious answer to a mysterious question[11]. What do you know, after you have said that intelligence is "emergent"? You can make no new predictions. You do not know anything about the behavior of real-world minds that you did not know before. It feels like you believe a new fact, but you don't anticipate any different outcomes. Your curiosity feels sated, but it has not been fed. The hypothesis has no moving parts—there's no detailed internal model to manipulate. Those who proffer the hypothesis of "emergence" confess their ignorance of the internals, and take pride in it; they contrast the science of "emergence" to other sciences merely mundane.

And even after the answer of "Why? Emergence!" is given, *the phenomenon is still a mystery* and possesses the same sacred impenetrability it had at the start.

A fun exercise is to eliminate the adjective "emergent" from any sentence in which it appears, and see if the sentence says anything different:

- *Before:* Human intelligence is an emergent product of neurons firing.
- *After:* Human intelligence is a product of neurons firing.

11. Page 96, 'Mysterious Answers to Mysterious Questions'.

- *Before:* The behavior of the ant colony is the emergent outcome of the interactions of many individual ants.
- *After:* The behavior of the ant colony is the outcome of the interactions of many individual ants.
- *Even better:* A colony is made of ants. We can successfully predict some aspects of colony behavior using models that include only individual ants, without any global colony variables, showing that we understand how those colony behaviors arise from ant behaviors.

Another fun exercise is to replace the word "emergent" with the old word[12], the explanation[13] that people had to use before emergence was invented:

- *Before:* Life is an emergent phenomenon.
- *After:* Life is a magical phenomenon.
- *Before:* Human intelligence is an emergent product of neurons firing.
- *After:* Human intelligence is a magical product of neurons firing.

Does not each statement convey exactly the same amount of knowledge about the phenomenon's behavior? Does not each hypothesis fit exactly the same set of outcomes[14]?

"Emergence" has become very popular, just as saying "magic" used to be very popular. "Emergence" has the same deep appeal to human psychology, for the same reason. "Emergence" is such a wonderfully easy explanation, and it feels good to say it; it gives you a sacred mystery[15] to worship. Emergence is popular *because* it is the junk food of curiosity. You can explain anything using emergence, and so people do just that; for it feels so wonderful to explain things. Humans are still humans, even if they've taken a few science classes in college. Once they

12. Page 80, 'Guessing the Teacher's Password'.
13. Page 77, 'Fake Explanations'.
14. Page 62, 'Your Strength as a Rationalist'.
15. Page 96, 'Mysterious Answers to Mysterious Questions'.

find a way to escape the shackles[16] of settled science, they get up to the same shenanigans as their ancestors, dressed up in the literary genre of "science"[17] but still the same species psychology.

16. *http://yudkowsky.net/virtues/*
17. Page 53, 'Belief as Attire'.

20. Say Not "Complexity"[1]

Once upon a time...

This is a story from when I first met Marcello, with whom I would later work for a year on AI theory; but at this point I had not yet accepted him as my apprentice. I knew that he competed at the national level in mathematical and computing olympiads, which sufficed to attract my attention for a closer look; but I didn't know yet if he could learn to think about AI.

I had asked Marcello to say how he thought an AI might discover how to solve a Rubik's Cube. Not in a preprogrammed way, which is trivial, but rather how the AI itself might figure out the laws of the Rubik universe and reason out how to exploit them. How would an AI *invent for itself* the concept of an "operator", or "macro", which is the key to solving the Rubik's Cube?

At some point in this discussion, Marcello said: "Well, I think the AI needs complexity to do X, and complexity to do Y—"

And I said, "Don't say '*complexity*'."

Marcello said, "Why not?"

I said, "Complexity should never be a goal in itself. You may need to use a particular algorithm that adds some amount of complexity, but complexity for the sake of complexity just makes things harder." (I was thinking of all the people whom I had heard advocating that the Internet would "wake up" and become an AI when it became "sufficiently complex".)

And Marcello said, "But there's got to be *some* amount of complexity that does it."

I closed my eyes briefly, and tried to think of how to explain it all in words. To me, saying 'complexity' simply *felt* like the wrong move in the AI dance. No one can think fast enough to deliberate, in words, about each sentence of their stream

1. *http://lesswrong.com/lw/ix/say_not_complexity/*

of consciousness; for that would require an infinite recursion. We think in words, but our stream of consciousness is steered below the level of words, by the trained-in remnants of past insights and harsh experience...

I said, "Did you read A Technical Explanation of Technical Explanation[2]?"

"Yes," said Marcello.

"Okay," I said, "saying 'complexity' doesn't concentrate your probability mass."

"Oh," Marcello said, "like 'emergence[3]'. Huh. So... now I've got to think about how X might actually happen..."

That was when I thought to myself, "*Maybe **this** one is teachable.*"

Complexity is not a useless concept. It has mathematical definitions attached to it, such as Kolmogorov complexity, and Vapnik-Chervonenkis complexity. Even on an intuitive level, complexity is often worth thinking about—you have to judge the complexity of a hypothesis and decide if it's "too complicated" given the supporting evidence, or look at a design and try to make it simpler.

But concepts are not useful or useless of themselves. Only *usages* are correct or incorrect. In the step Marcello was trying to take in the dance, he was trying to explain something for free, get something for nothing. It is an extremely common misstep, at least in my field. You can join a discussion on Artificial General Intelligence and watch people doing the same thing, left and right, over and over again—constantly skipping over things they don't understand, without realizing that's what they're doing.

In an eyeblink it happens: putting a non-controlling causal node[4] behind something mysterious, a causal node that feels

2. *http://yudkowsky.net/bayes/technical.html*
3. Page 100, 'The Futility of Emergence'.
4. Page 87, 'Fake Causality'.

like an explanation[5] but isn't. The mistake takes place below the level of words. It requires no special character flaw; it is how human beings think by default[6], since the ancient times.

What you must avoid is *skipping over the mysterious part;* you must linger at the mystery to confront it directly. There are many words that can skip over mysteries, and some of them would be legitimate in other contexts—"complexity", for example. But the essential mistake is that *skip-over,* regardless of what causal node goes behind it. The skip-over is not a thought, but a microthought. You have to pay close attention to catch yourself at it. And when you train yourself to avoid skipping, it will become a matter of instinct, not verbal reasoning. You have to *feel* which parts of your map are still blank, and more importantly, pay attention to that feeling.

I suspect that in academia there is a huge pressure to sweep problems under the rug so that you can present a paper with the appearance of completeness. You'll get more kudos for a seemingly complete model that includes some "emergent phenomena[7]", versus an explicitly incomplete map where the label says "I got no clue how this part works" or "then a miracle occurs". A journal may not even accept the latter paper, since who knows but that the unknown steps are really where everything interesting happens? And yes, it sometimes happens that all the non-magical parts of your map turn out to also be non-important. That's the price you sometimes pay, for entering into terra incognita and trying to solve problems *incrementally*. But that makes it even *more* important to *know* when you aren't finished yet. Mostly, people don't dare to enter terra incognita at all, for the deadly fear of wasting their time.

And if you're working on a revolutionary AI startup, there is an even huger pressure to sweep problems under the rug; or you will have to admit to yourself[8] that you don't know how to

5. Page 77, 'Fake Explanations'.
6. Page 80, 'Guessing the Teacher's Password'.
7. Page 100, 'The Futility of Emergence'.

build an AI yet, and your current life-plans will come crashing down in ruins around your ears. But perhaps I am over-explaining[9], since skip-over happens by default[10] in humans; if you're looking for examples, just watch people discussing religion or philosophy or spirituality or any science in which they were not professionally trained.

Marcello and I developed a convention in our AI work: when we ran into something we didn't understand, which was often, we would say "magic"—as in, "X magically does Y"—to remind ourselves that *here was an unsolved problem, a gap in our understanding*. It is far better to say "magic", than "complexity" or "emergence"; the latter words[11] create an illusion of understanding. Wiser to say "magic", and leave yourself a placeholder, a reminder of work you will have to do later.

8. Page 477, 'You Can Face Reality'.
9. Page 157, 'Correspondence Bias'.
10. Page 80, 'Guessing the Teacher's Password'.
11. Page 80, 'Guessing the Teacher's Password'.

21. Positive Bias: Look Into the Dark[1]

I am teaching a class, and I write upon the blackboard three numbers: 2-4-6. "I am thinking of a rule," I say, "which governs sequences of three numbers. The sequence 2-4-6, as it so happens, obeys this rule. Each of you will find, on your desk, a pile of index cards. Write down a sequence of three numbers on a card, and I'll mark it "Yes" for fits the rule, or "No" for not fitting the rule. Then you can write down another set of three numbers and ask whether it fits again, and so on. When you're confident that you know the rule, write down the rule on a card. You can test as many triplets as you like."

Here's the record of one student's guesses:

4, 6, 2	No
4, 6, 8	Yes
10, 12, 14	Yes

At this point the student wrote down his guess at the rule. What do *you* think the rule is? Would you have wanted to test another triplet, and if so, what would it be? Take a moment to think before continuing.

The challenge above is based on a classic experiment due to Peter Wason, the 2-4-6 task. Although subjects given this task typically expressed high confidence in their guesses, only 21% of the subjects successfully guessed the experimenter's real rule, and replications since then have continued to show success rates of around 20%.

The study was called "On the failure to eliminate hypotheses in a conceptual task" (*Quarterly Journal of Experimental Psychology,* 12: 129-140, 1960). Subjects who attempt the 2-4-6 task usually try to generate *positive* examples, rather than *negative* examples—they apply the hypothetical rule to generate a representative instance, and see if it is labeled "Yes".

1. *http://lesswrong.com/lw/iw/positive_bias_look_into_the_dark/*

Thus, someone who forms the hypothesis "numbers increasing by two" will test the triplet 8-10-12, hear that it fits, and confidently announce the rule. Someone who forms the hypothesis X-2X-3X will test the triplet 3-6-9, discover that it fits, and then announce that rule.

In every case the actual rule is the same: the three numbers must be in ascending order.

But to discover this, you would have to generate triplets that *shouldn't* fit, such as 20-23-26, and see if they are labeled "No". Which people tend not to do, in this experiment. In some cases, subjects devise, "test", and announce rules far more complicated than the actual answer.

This cognitive phenomenon is usually lumped in with "confirmation bias". However, it seems to me that the phenomenon of trying to test *positive* rather than *negative* examples, ought to be distinguished from the phenomenon of trying to preserve the belief you started with. "Positive bias" is sometimes used as a synonym for "confirmation bias", and fits this particular flaw much better.

It once seemed that phlogiston theory[2] could explain a flame going out in an enclosed box (the air became saturated with phlogiston and no more could be released), but phlogiston theory could just as well have explained the flame *not* going out. To notice this, you have to search for negative examples instead of positive examples, look into zero instead of one; which goes against the grain of what experiment has shown to be human instinct.

For by instinct, we human beings only live in half the world.

One may be lectured on positive bias for days, and yet overlook it in-the-moment. Positive bias is not something we do as a matter of logic, or even as a matter of emotional attachment. The 2-4-6 task is "cold", logical, not affectively "hot". And yet the mistake is sub-verbal, on the level of imagery, of instinctive reactions. Because the problem doesn't arise from following a

2. Page 87, 'Fake Causality'.

deliberate rule that says "Only think about positive examples", it can't be solved just by knowing verbally that "We ought to think about both positive and negative examples." Which example automatically pops into your head? You have to learn, wordlessly, to zag instead of zig. You have to learn to flinch toward the zero, instead of away from it.

I have been writing for quite some time now on the notion that the strength of a hypothesis is what it *can't* explain, not what it *can*[3]—if you are equally good at explaining any outcome, you have zero knowledge. So to spot an explanation that isn't helpful, it's not enough to think of what it does explain very well—you also have to search for results it *couldn't* explain, and this is the true strength of the theory.

So I said all this, and then yesterday, I challenged the usefulness of "emergence" as a concept[4]. One commenter cited superconductivity and ferromagnetism as examples of emergence. I replied that non-superconductivity and non-ferromagnetism were also examples of emergence, which was the problem. But be it far from me to criticize the commenter! Despite having read extensively on "confirmation bias", I didn't spot the "gotcha" in the 2-4-6 task the first time I read about it. It's a subverbal blink-reaction that has to be retrained. I'm still working on it myself.

So much of a rationalist's skill is below the level of words. It makes for challenging work in trying to convey the Art through blog posts. People will agree with you, but then, in the next sentence, do something subdeliberative that goes in the opposite direction. Not that I'm complaining! A major reason I'm posting here is to observe what my words *haven't* conveyed.

Are you searching for positive examples of positive bias right now, or sparing a fraction of your search on what positive bias should lead you to *not* see? Did you look toward light or darkness?

3. Page 62, 'Your Strength as a Rationalist'.
4. Page 100, 'The Futility of Emergence'.

22. My Wild and Reckless Youth[1]

It is said that parents do all the things they tell their children not to do, which is how they know not to do them.

Long ago, in the unthinkably distant past, I was a devoted Traditional Rationalist, conceiving myself skilled according to that kind, yet I knew not the Way of Bayes. When the young Eliezer was confronted with a mysterious-seeming question, the precepts of Traditional Rationality did not stop him from devising a Mysterious Answer[2]. It is, by far, the most embarrassing mistake I made in my life, and I still wince to think of it.

What was my mysterious answer to a mysterious question? This I will not describe, for it would be a long tale and complicated. I was young, and a mere Traditional Rationalist who knew not the teachings of Tversky and Kahneman. I knew about Occam's Razor, but not the conjunction fallacy[3]. I thought I could get away with thinking complicated thoughts myself, in the literary style of the complicated thoughts I read in science books, not realizing that correct complexity is only possible when every step is pinned down overwhelmingly. Today, one of the chief pieces of advice I give to aspiring young rationalists is "Do not attempt long chains of reasoning or complicated plans."

Nothing more than this need be said: Even after I invented my "answer", the phenomenon was still a mystery[4] unto me, and possessed the same quality of wondrous impenetrability that it had at the start.

Make no mistake[5], that younger Eliezer was not stupid. All the errors of which the young Eliezer was guilty, are still being made today by respected scientists in respected journals. It

1. *http://lesswrong.com/lw/iy/my_wild_and_reckless_youth/*
2. Page 96, 'Mysterious Answers to Mysterious Questions'.
3. *http://en.wikipedia.org/wiki/Conjunction_fallacy*
4. Page 96, 'Mysterious Answers to Mysterious Questions'.
5. Page 157, 'Correspondence Bias'.

would have taken a subtler skill to protect him, than ever he was taught as a Traditional Rationalist.

Indeed, the young Eliezer diligently and painstakingly followed the injunctions of Traditional Rationality in the course of going astray.

As a Traditional Rationalist, the young Eliezer was careful to ensure that his Mysterious Answer made a bold prediction of future experience. Namely, I expected future neurologists to discover that neurons were exploiting quantum gravity, a la Sir Roger Penrose. This required neurons to maintain a certain degree of quantum coherence, which was something you could look for, and find or not find. Either you observe that or you don't, right?

But my hypothesis made no *retrospective* predictions. According to Traditional Science, retrospective predictions don't count—so why bother making them? To a Bayesian, on the other hand, if a hypothesis does not *today* have a favorable likelihood ratio over "I don't know", it raises the question of why you *today* believe anything more complicated than "I don't know". But I knew not the Way of Bayes, so I was not thinking about likelihood ratios or focusing probability density. I had Made a Falsifiable Prediction; was this not the Law?

As a Traditional Rationalist, the young Eliezer was careful not to believe in magic, mysticism, carbon chauvinism, or anything of that sort. I proudly professed[6] of my Mysterious Answer, "It is just physics like all the rest of physics!" As if you could save magic from being a cognitive isomorph of magic, by calling[7] it quantum gravity. But I knew not the Way of Bayes, and did not see the level[8] on which my idea was isomorphic to magic. I gave my *allegiance* to physics, but this did not save me; what does probability theory know of allegiances? I

6. Page 50, 'Professing and Cheering'.
7. Page 84, 'Science as Attire'.
8. Page 77, 'Fake Explanations'.

avoided everything that Traditional Rationality told me was forbidden, but what was left was still magic.

Beyond a doubt, my allegiance to Traditional Rationality helped me get out of the hole I dug myself into. If I hadn't been a Traditional Rationalist, I would have been *completely* screwed. But Traditional Rationality still wasn't enough to get it *right*. It just led me into different mistakes than the ones it had explicitly forbidden.

When I think about how my younger self very carefully followed the rules of Traditional Rationality in the course of getting the answer *wrong,* it sheds light on the question of why people who call themselves "rationalists" do not rule the world[9]. You need *one whole hell of a lot* of rationality before it does anything but lead you into new and interesting mistakes.

Traditional Rationality is taught as an art, rather than a science; you read the biography of famous physicists describing the lessons life taught them, and you try to do what they tell you to do. But you haven't lived their lives, and half of what they're trying to describe is an instinct that has been trained into them.

The way Traditional Rationality is designed, it would have been acceptable for me to spend 30 years on my silly idea, so long as I succeeded in falsifying it eventually, and was honest with myself about what my theory predicted, and accepted the disproof when it arrived, et cetera. This is enough to let the Ratchet of Science click forward, but it's a little harsh on the people who waste 30 years of their lives. Traditional Rationality is a walk, not a dance. It's designed to get you to the truth *eventually*, and gives you all too much time to smell the flowers along the way.

Traditional Rationalists can agree to disagree. Traditional Rationality doesn't have the *ideal* that thinking is an exact art in which there is only one correct probability estimate given the evidence. In Traditional Rationality, you're allowed to guess,

9. Page 333, 'Knowing About Biases Can Hurt People'.

and then test your guess. But experience has taught me that if you don't *know,* and you guess, you'll end up being wrong.

The Way of Bayes is also an imprecise art, at least the way I'm holding forth upon it. These blog posts are still fumbling attempts to put into words lessons that would be better taught by experience. But at least there's *underlying* math, plus experimental evidence from cognitive psychology on how humans actually think. Maybe that will be enough to cross the stratospherically high threshold required for a discipline that lets you actually get it right, instead of just constraining you into interesting new mistakes.

23. Failing to Learn from History[1]

Continuation of: My Wild and Reckless Youth[2]

Once upon a time, in my wild and reckless youth[3], when I knew not the Way of Bayes, I gave a Mysterious Answer[4] to a mysterious-seeming question. Many failures occurred in sequence, but one mistake stands out as most critical: My younger self did not realize that *solving a mystery should make it feel less confusing*. I was trying to explain a Mysterious Phenomenon—which to me meant providing a cause for it, fitting it into an integrated model of reality. Why should this make the phenomenon less Mysterious, when that is its nature? I was trying to *explain* the Mysterious Phenomenon, not render it (by some impossible alchemy) into a mundane phenomenon, a phenomenon that wouldn't even call out for an unusual explanation in the first place.

As a Traditional Rationalist, I knew the historical tales of astrologers and astronomy, of alchemists and chemistry, of vitalists and biology. But the Mysterious Phenomenon was not like this. It was something *new,* something stranger, something more difficult, something that ordinary science had failed to explain for centuries—

- as if stars and matter and life had not been mysteries for hundreds of years and thousands of years, from the dawn of human thought right up until science finally solved them—

We learn about astronomy and chemistry and biology in school, and it seems to us that these matters have *always been* the proper realm of science, that they have *never been* mysterious. When science dares to challenge a new Great Puzzle, the children of that generation are skeptical, for they have never

1. *http://lesswrong.com/lw/iz/failing_to_learn_from_history/*
2. Page 112, 'My Wild and Reckless Youth'.
3. Page 112, 'My Wild and Reckless Youth'.
4. Page 96, 'Mysterious Answers to Mysterious Questions'.

seen science explain something that *feels* mysterious to them. Science is only good for explaining *scientific* subjects, like stars and matter and life.

I thought the lesson of history was that astrologers and alchemists and vitalists had an innate character flaw[5], a tendency toward mysterianism, which led them to come up with mysterious explanations for non-mysterious subjects. But surely, if a phenomenon really *was* very weird, a weird explanation might be in order?

It was only afterward, when I began to see the mundane structure inside the mystery, that I realized whose shoes I was standing in. Only then did I realize how reasonable vitalism had seemed *at the time*, how *surprising* and *embarrassing* had been the universe's reply of, "Life is mundane, and does not need a weird explanation."

We read history but we don't *live* it, we don't *experience* it. If only I had *personally* postulated astrological mysteries and then discovered Newtonian mechanics, postulated alchemical mysteries and then discovered chemistry, postulated vitalistic mysteries and then discovered biology. I would have thought of my Mysterious Answer and said to myself: *No way am I falling for that again.*

5. Page 157, 'Correspondence Bias'.

24. Making History Available[1]

Followup to: Failing to Learn from History[2]

There is a habit of thought which I call the *logical fallacy of generalization from fictional evidence,* which deserves a blog post in its own right, one of these days. Journalists who, for example, talk about the *Terminator* movies in a report on AI, do not usually treat *Terminator* as a prophecy or fixed truth. But the movie is recalled—is available[3]—as if it were an illustrative historical case. As if the journalist had seen it happen on some other planet, so that it might well happen here. More on this in Section 6 of this paper[4].

There is an inverse error to generalizing from fictional evidence: failing to be sufficiently moved by *historical* evidence. The trouble with generalizing from fictional evidence is that it is fiction—it never actually happened. It's not drawn from the same distribution as this, our real universe; fiction differs from reality in systematic ways[5]. But history *has* happened, and *should* be available.

In our ancestral environment, there were no movies; what you saw with your own eyes was true. Is it any wonder that fictions we see in lifelike moving pictures have too great an impact on us? Conversely, things that *really happened,* we encounter as ink on paper; they happened, but we never *saw* them happen. We don't remember them happening to us.

The inverse error is to treat history as mere story, process it with the same part of your mind that handles the novels you read. You may say with your lips that it is "truth", rather than "fiction", but that doesn't mean you are being moved as much as

1. *http://lesswrong.com/lw/jo/making_history_available/*
2. Page 116, 'Failing to Learn from History'.
3. *http://en.wikipedia.org/wiki/Availability_heuristic*
4. *http://intelligence.org/Biases.pdf*
5. *http://www.overcomingbias.com/2007/07/tell-your-anti-.html*

you should be. Many biases involve being insufficiently moved by dry, abstract information[6].

Once upon a time, I gave a Mysterious Answer[7] to a mysterious question, not realizing that I was making exactly the same mistake as astrologers devising mystical explanations for the stars, or alchemists devising magical properties of matter, or vitalists postulating an opaque "elan vital" to explain all of biology.

When I finally realized whose shoes I was standing in[8], there was a sudden shock of unexpected connection with the past. I realized that the invention and destruction of vitalism—which I had only read about in books—had *actually happened to real people,* who experienced it much the same way I experienced the invention and destruction of my own mysterious answer. And I also realized that if I had actually *experienced* the past—if I had lived through past scientific revolutions myself, rather than reading about them in history books—I probably would *not* have made the same mistake again. I would not have come up with *another* mysterious answer; the first thousand lessons would have hammered home the moral.

So (I thought), to feel sufficiently the force of history, I should try to approximate the thoughts of an Eliezer who *had* lived through history—I should try to think as if everything I read about in history books, had actually happened to me. (With appropriate reweighting for the availability bias of history books—I should remember being a thousand peasants for every ruler.) I should immerse myself in history, imagine *living* through eras I only saw as ink on paper.

Why should I remember the Wright Brothers' first flight? I was not there. But as a rationalist, could I dare to *not* remember, when the event actually happened? Is there so much difference between seeing an event through your eyes—which is

6. *http://lesswrong.com/lw/hw/scope_insensitivity/*
7. Page 96, 'Mysterious Answers to Mysterious Questions'.
8. Page 116, 'Failing to Learn from History'.

actually a causal chain involving reflected photons, not a direct connection—and seeing an event through a history book? Photons and history books both descend by causal chains from the event itself.

I had to overcome the false amnesia of being born at a particular time. I had to recall—make available[9]—*all* the memories, not just the memories which, by mere coincidence, belonged to myself and my own era.

The Earth became older, of a sudden.

To my former memory, the United States had always existed—there was never a time when there was no United States. I had not remembered, until that time, how the Roman Empire rose, and brought peace and order, and lasted through so many centuries, until I forgot that things had ever been otherwise; and yet the Empire fell, and barbarians overran my city, and the learning that I had possessed was lost. The modern world became more fragile to my eyes; it was not the first modern world.

So many mistakes, made over and over and *over* again, because I did not remember making them, in every era I never lived...

And to think, people sometimes wonder if overcoming bias is important.

Don't you remember how many times your biases have killed you? You don't? I've noticed that sudden amnesia often follows a fatal mistake. But take it from me, it happened. I remember; I wasn't there.

So the next time you doubt the strangeness of the future, remember how you were born in a hunter-gatherer tribe ten thousand years ago, when no one knew of Science at all. Remember how you were shocked, to the depths of your being, when Science explained the great and terrible sacred mysteries that you once revered so highly. Remember how you once believed that you could fly by eating the right mushrooms, and then you accepted with disappointment that you would never fly, and then

9. *http://en.wikipedia.org/wiki/Availability_heuristic*

you flew. Remember how you had always thought that slavery was right and proper, and then you changed your mind. Don't imagine how you *could* have predicted the change[10], for that is amnesia. *Remember* that, in fact, you did not guess. Remember how, century after century, the world changed in ways you did not guess.

Maybe then you will be less shocked by what happens next.

10. Page 71, 'Hindsight bias'.

25. Explain/Worship/Ignore?[1]

Followup to: Semantic Stopsigns[2], Mysterious Answers to Mysterious Questions[3]

As our tribe wanders through the grasslands, searching for fruit trees and prey, it happens every now and then that water pours down from the sky.

"Why does water sometimes fall from the sky?" I ask the bearded wise man of our tribe.

He thinks for a moment, this question having never occurred to him before, and then says, "From time to time, the sky spirits battle, and when they do, their blood drips from the sky."

"Where do the sky spirits come from?" I ask.

His voice drops to a whisper. "From the before time. From the long long ago."

When it rains, and you don't know why, you have several options. First, you could simply not ask why—not follow up on the question, or never think of the question in the first place. This is the Ignore command, which the bearded wise man originally selected. Second, you could try to devise some sort of explanation, the Explain command, as the bearded man did in response to your first question. Third, you could enjoy the sensation of mysteriousness—the Worship command.

Now, as you are bound to notice from this story, each time you select Explain, the best-case scenario is that you get an explanation, such as "sky spirits". But then this explanation itself is subject to the same dilemma—Explain, Worship, or Ignore? Each time you hit Explain, science grinds for a while, returns an explanation, and then another dialog box pops up. As good rationalists, we feel duty-bound to keep hitting Explain, but it seems like a road that has no end.

---

1. *http://lesswrong.com/lw/j2/explainworshipignore/*
2. Page 92, 'Semantic Stopsigns'.
3. Page 96, 'Mysterious Answers to Mysterious Questions'.

You hit Explain for life, and get chemistry; you hit Explain for chemistry, and get atoms; you hit Explain for atoms, and get electrons and nuclei; you hit Explain for nuclei, and get quantum chromodynamics and quarks; you hit Explain for how the quarks got there, and get back the Big Bang...

We can hit Explain for the Big Bang, and wait while science grinds through its process, and maybe someday it will return a perfectly good explanation. But then that will just bring up another dialog box. So, if we continue long enough, we must come to a *special* dialog box, a *new* option, an Explanation That Needs No Explanation, a place where the chain ends—and this, maybe, is the only explanation worth knowing.

There—I just hit Worship.

Never forget that there are many more ways to worship something than lighting candles around an altar.

If I'd said, "Huh, that does seem paradoxical. I wonder how the apparent paradox is resolved?" then I would have hit Explain, which does sometimes take a while to produce an answer.

And if the whole issue seems to you unimportant, or irrelevant, or if you'd rather put off thinking about it until tomorrow, than you have hit Ignore.

Select your option wisely.

26. "Science" as Curiosity-Stopper[1]

Followup to: Semantic Stopsigns[2], Mysterious Answers to Mysterious Questions[3], Say Not 'Complexity'[4]

Imagine that I, in full view of live television cameras, raised my hands and chanted *abracadabra* and caused a brilliant light to be born, flaring in empty space beyond my outstretched hands. Imagine that I committed this act of blatant, unmistakeable sorcery under the full supervision of James Randi and all skeptical armies. Most people, I think, would be *fairly curious* as to what was going on.

But now suppose instead that I don't go on television. I do not wish to share the power, nor the truth behind it. I want to keep my sorcery secret. And yet I also want to cast my spells whenever and wherever I please. I want to cast my brilliant flare of light so that I can read a book on the train—without anyone becoming curious. Is there a spell that stops curiosity?

Yes indeed! Whenever anyone asks "How did you do that?", I just say "Science!"

It's not a real explanation[5], so much as a curiosity-stopper[6]. It doesn't tell you whether the light will brighten or fade, change color in hue or saturation, and it certainly doesn't tell you how to make a similar light yourself. You don't actually *know* anything more than you knew before I said the magic word[7]. But you turn away, satisfied that nothing unusual is going on.

Better yet, the same trick works with a standard light switch.

Flip a switch and a light bulb turns on. Why?

---

1. *http://lesswrong.com/lw/j3/science_as_curiositystopper/*
2. Page 92, 'Semantic Stopsigns'.
3. Page 96, 'Mysterious Answers to Mysterious Questions'.
4. Page 104, 'Say Not "Complexity"'.
5. Page 77, 'Fake Explanations'.
6. Page 92, 'Semantic Stopsigns'.
7. Page 80, 'Guessing the Teacher's Password'.

In school, one is taught that the password[8] to the light bulb is "Electricity!" By now, I hope, you're wary of marking the light bulb "understood" on such a basis. Does saying "Electricity!" let you do calculations that will control your anticipation of experience? There is, at the least, a great deal more to learn. (Physicists should ignore this paragraph and substitute a problem in evolutionary theory[9], where the substance of the theory is again in calculations that few people know how to perform.)

If you thought the light bulb was *scientifically inexplicable,* it would seize the *entirety* of your attention. You would drop whatever else you were doing, and focus on that light bulb.

But what does the phrase "scientifically explicable" mean? It means that someone *else* knows how the light bulb works. When you are told the light bulb is "scientifically explicable", you don't know more than you knew earlier; you don't know whether the light bulb will brighten or fade. But because someone *else* knows, it devalues the knowledge in your eyes. You become less curious.

Since this is an econblog, someone out there is bound to say, "If the light bulb were unknown to science, you could gain fame and fortune by investigating it." But I'm not talking about greed. I'm not talking about career ambition. I'm talking about the raw emotion of curiosity—the feeling of being intrigued. Why should *your* curiosity be diminished because someone *else,* not you, knows how the light bulb works? Is this not spite? It's not enough for *you* to know; other people must also be ignorant, or you won't be happy?

There are goods that knowledge may serve besides curiosity, such as the social utility of technology. For these instrumental goods, it matters whether some other entity in local space already knows. But for my own curiosity, why should it matter?

Besides, consider the consequences if you permit "Someone else knows the answer" to function as a curiosity-stopper. One

8. Page 80, 'Guessing the Teacher's Password'.
9. *http://lesswrong.com/lw/kr/an_alien_god/*

day you walk into your living room and see a giant green elephant, seemingly hovering in midair, surrounded by an aura of silver light.

"What the heck?" you say.

And a voice comes from above the elephant, saying, "SOMEBODY ALREADY KNOWS WHY THIS ELEPHANT IS HERE[10]."

"Oh," you say, "in that case, never mind," and walk on to the kitchen.

I don't know the grand unified theory for this universe's laws of physics. I also don't know much about human anatomy with the exception of the brain. I couldn't point out on my body where my kidneys are, and I can't recall offhand what my liver does. (I am not proud of this. Alas, with all the math I need to study, I'm not likely to learn anatomy anytime soon.)

Should I, so far as *curiosity* is concerned, be more intrigued by my ignorance of the ultimate laws of physics, than the fact that I don't know much about what goes on inside my own body?

If I raised my hands and cast a light spell, you would be intrigued. Should you be any *less* intrigued by the very fact that I raised my hands? When you raise your arm and wave a hand around, this act of will is coordinated by (among other brain areas) your cerebellum. I bet you don't know how the cerebellum works. *I* know a little—though only the gross details, not enough to perform calculations... but so what? What does that matter, if *you* don't know? Why should there be a double standard of curiosity for sorcery and hand motions?

Look at yourself in the mirror. Do you know what you're looking at? Do you know what looks out from behind your eyes? Do you know what you are? Some of that answer, Science knows, and some of it Science does not. But why should that distinction matter to your curiosity, if *you* don't know?

10. *http://godescalc.wordpress.com/2012/06/24/overlooked-elephant/*

Do you know how your knees work? Do you know how your shoes were made? Do you know why your computer monitor glows? Do you know why water is wet?

The world around you is full of puzzles. Prioritize, if you must. But do not complain that cruel Science has emptied the world of mystery. With reasoning such as that, I could get you to overlook an elephant in your living room.

27. Applause Lights[1]

Followup to: Semantic Stopsigns[2], We Don't Really Want Your Participation[3]

At the Singularity Summit 2007, one of the speakers called for democratic, multinational development of AI. So I stepped up to the microphone and asked:

> Suppose that a group of democratic republics form a consortium to develop AI, and there's a lot of politicking during the process—some interest groups have unusually large influence, others get shafted—in other words, the result looks just like the products of modern democracies. Alternatively, suppose a group of rebel nerds develops an AI in their basement, and instructs the AI to poll everyone in the world—dropping cellphones to anyone who doesn't have them—and do whatever the majority says. Which of these do you think is more "democratic", and would you feel safe with either?

I wanted to find out whether he believed in the pragmatic adequacy of the democratic political process, or if he believed in the moral rightness of voting. But the speaker replied:

> The first scenario sounds like an editorial in Reason magazine, and the second sounds like a Hollywood movie plot.

Confused, I asked:

> Then what kind of democratic process *did* you have in mind?

1. *http://lesswrong.com/lw/jb/applause_lights/*
2. Page 92, 'Semantic Stopsigns'.
3. *http://lesswrong.com/lw/ja/we_dont_really_want_your_participation/*

The speaker replied:

> Something like the Human Genome Project—that was an internationally sponsored research project.

I asked:

> How would different interest groups resolve their conflicts in a structure like the Human Genome Project?

And the speaker said:

> I don't know.

This exchange puts me in mind of a quote[4] (~~which I failed to Google~~ found by Jeff Grey and Miguel) from some dictator or other, who was asked if he had any intentions to move his pet state toward democracy:

> We believe we are already within a democratic system. Some factors are still missing, like the expression of the people's will.

The substance of a democracy is the specific mechanism that resolves policy conflicts. If all groups had the same preferred policies, there would be no need for democracy—we would automatically cooperate. The resolution process can be a direct majority vote, or an elected legislature, or even a voter-sensitive behavior of an AI, but it has to be *something*. What does it *mean* to call for a "democratic" solution if you don't have a conflict-resolution mechanism in mind?

I think it means that you have said the word "democracy", so the audience is supposed to cheer. It's not so much a *proposi-*

4. *http://www.time.com/time/magazine/article/0,9171,954853,00.html*

tional statement, as the equivalent of the "Applause" light that tells a studio audience when to clap.

This case is remarkable only in that I mistook the applause light for a policy suggestion, with subsequent embarrassment for all. Most applause lights are much more blatant, and can be detected by a simple reversal test. For example, suppose someone says:

> We need to balance the risks and opportunities of AI.

If you reverse this statement, you get:

> We shouldn't balance the risks and opportunities of AI.

Since the reversal sounds *ab*normal, the unreversed statement is probably normal, implying it does not convey new information. There are plenty of legitimate reasons for uttering a sentence that would be uninformative in isolation. "We need to balance the risks and opportunities of AI" can introduce a discussion topic; it can emphasize the importance of a specific proposal for balancing; it can criticize an unbalanced proposal. Linking to a normal assertion can convey new information to a bounded rationalist—the link itself may not be obvious. But if *no* specifics follow, the sentence is probably an applause light.

I am tempted to give a talk sometime that consists of *nothing but* applause lights, and see how long it takes for the audience to start laughing:

> I am here to propose to you today that we need to balance the risks and opportunities of advanced Artificial Intelligence. We should avoid the risks and, insofar as it is possible, realize the opportunities. We should not needlessly confront entirely unnecessary dangers. To achieve these goals, we must plan wisely and rationally. We should not act in fear and panic,

or give in to technophobia; but neither should we act in blind enthusiasm. We should respect the interests of all parties with a stake in the Singularity. We must try to ensure that the benefits of advanced technologies accrue to as many individuals as possible, rather than being restricted to a few. We must try to avoid, as much as possible, violent conflicts using these technologies; and we must prevent massive destructive capability from falling into the hands of individuals. We should think through these issues before, not after, it is too late to do anything about them...

28. Truly Part Of You[1]

Followup to: Guessing the Teacher's Password[2], Artificial Addition[3]

A classic paper by Drew McDermott, "Artificial Intelligence Meets Natural Stupidity[4]", criticized AI programs that would try to represent notions like *happiness is a state of mind* using a semantic network:

```
STATE-OF-MIND
     ^
     | IS-A
     |
HAPPINESS
```

And of course there's nothing *inside* the "HAPPINESS" node; it's just a naked LISP token with a suggestive English name.

So, McDermott says, "A good test for the disciplined programmer is to try using gensyms in key places and see if he still admires his system. For example, if STATE-OF-MIND is renamed G1073..." then we would have `IS-A(HAPPINESS, G1073)` "which looks much more dubious."

Or as I would slightly rephrase the idea: If you substituted randomized symbols for *all* the suggestive English names, you would be completely unable to figure out what `G1071(G1072, 1073)` meant. Was the AI program meant to represent hamburgers? Apples? Happiness? Who knows? *If you delete the suggestive English names, they don't grow back.*

1. *http://lesswrong.com/lw/la/truly_part_of_you/*
2. Page 80, 'Guessing the Teacher's Password'.
3. *http://lesswrong.com/lw/l9/artificial_addition/*
4. *http://ontolog.cim3.net/forum/ontolog-forum/2006-04/pdfb9P8wB8aYh.pdf*

Suppose a physicist tells you that "Light is waves[5]", and you *believe* him. You now have a little network in your head that says IS-A(LIGHT, WAVES). If someone asks you "What is light made of?" you'll be able to say "Waves!"

As McDermott says, "The whole problem is getting the hearer to notice what it has been told. Not 'understand', but 'notice'." Suppose that instead the physicist told you, "Light is made of little curvy things." (Not true, btw.) Would you *notice* any difference of anticipated experience[6]?

How can you realize that you shouldn't trust your seeming knowledge that "light is waves"? One test you could apply is asking, "Could I *regenerate* this knowledge if it were somehow deleted from my mind?"

This is similar in spirit to scrambling the names of suggestively named LISP tokens in your AI program, and seeing if someone else can figure out what they allegedly "refer" to. It's also similar in spirit to observing that while an Artificial Arithmetician[7] can record and play back `Plus-Of(Seven, Six) = Thirteen`, it can't regenerate the knowledge if you delete it from memory, until another human re-enters it in the database. Just as if you forgot that "light is waves", you couldn't get back the knowledge except the same way you got the knowledge to begin with—by asking a physicist. You couldn't generate the knowledge for yourself, the way that physicists originally generated it.

The same experiences that lead us to formulate a belief, connect that belief to other knowledge and sensory input and motor output. If you see a beaver chewing a log, then you know what this thing-that-chews-through-logs looks like, and you will be able to recognize it on future occasions whether it is called a "beaver" or not. But if you acquire your beliefs about

5. Page 80, 'Guessing the Teacher's Password'.
6. Page 39, 'Making Beliefs Pay Rent (in Anticipated Experiences)'.
7. *http://lesswrong.com/lw/l9/artificial_addition/*

beavers by someone else telling you facts about "beavers", you may not be able to recognize a beaver when you see one.

This is the terrible danger of trying to *tell* an Artificial Intelligence facts which it could not learn for itself. It is also the terrible danger of trying to *tell* someone about physics that they cannot verify for themselves. For what physicists mean by "wave" is not "little squiggly thing" but a purely mathematical concept.

As Davidson observes, if you believe that "beavers" live in deserts, are pure white in color, and weigh 300 pounds when adult, then you do not have any beliefs *about* beavers, true or false. Your belief about "beavers" is not right enough to be wrong. If you don't have enough experience to regenerate beliefs when they are deleted, then do you have enough experience to connect that belief to anything at all? Wittgenstein: "A wheel that can be turned though nothing turns with it, is not part of the mechanism."

Almost as soon as I started reading about AI—even before I read McDermott—I realized it would be *a really good idea* to always ask myself: "How would I regenerate this knowledge if it were deleted from my mind?"

The deeper the deletion, the stricter the test. If all proofs of the Pythagorean Theorem were deleted from my mind, could I re-prove it? I think so. If all knowledge of the Pythagorean Theorem were deleted from my mind, would I notice the Pythagorean Theorem to re-prove? That's harder to boast, without putting it to the test; but if you handed me a right triangle with sides 3 and 4, and told me that the length of the hypotenuse was calculable, I think I would be able to calculate it, if I still knew all the rest of my math.

What about the notion of *mathematical proof?* If no one had ever told it to me, would I be able to reinvent *that* on the basis of other beliefs I possess? There was a time when humanity did not have such a concept. Someone must have invented it. What was it that they noticed? Would I notice if I saw some-

thing equally novel and equally important? Would I be able to think that far outside the box[8]?

How much of your knowledge could you regenerate? From how deep a deletion? It's not just a test to cast out insufficiently connected beliefs. It's a way of absorbing *a fountain of knowledge, not just one fact.*

A shepherd builds a counting system[9] that works by throwing a pebble into a bucket whenever a sheep leaves the fold, and taking a pebble out whenever a sheep returns. If you, the apprentice, do not understand this system—if it is magic that works for no apparent reason—then you will not know what to do if you accidentally drop an extra pebble into the bucket. That which you cannot make yourself, you cannot *remake* when the situation calls for it. You cannot go back to the source, tweak one of the parameter settings, and regenerate the output, without the source. If "Two plus four equals six" is a brute fact unto you, and then one of the elements changes to "five", how are you to know that "two plus five equals seven" when you were simply *told* that "two plus four equals six"?

If you see a small plant that drops a seed whenever a bird passes it, it will not occur to you that you can use this plant to partially automate the sheep-counter. Though you learned something that the original maker would use to improve on his invention, you can't go back to the source and re-create it.

When you contain the source of a thought, that thought can change along with you as you acquire new knowledge and new skills. When you contain the source of a thought, it becomes truly a part of you and grows along with you.

Strive to make yourself the source of every thought worth thinking. If the thought originally came from outside, make sure it comes from inside as well. Continually ask yourself: "How would I regenerate the thought if it were deleted?" When

8. Page 300, 'The "Outside the Box" Box'.
9. *http://sl4.org/wiki/TheSimpleTruth*

you have an answer, imagine *that* knowledge being deleted as well. And when you find a fountain, see what else it can pour.

29. Chaotic Inversion[1]

I was recently having a conversation with some friends on the topic of hour-by-hour productivity and willpower maintenance—something I've struggled with my whole life.

I can avoid running away from a hard problem the first time I see it[2] (perseverance on a timescale of seconds), and I can stick to the same problem for years; but to keep working on a timescale of *hours* is a constant battle for me. It goes without saying that I've already read reams and reams of advice; and the most help I got from it was realizing that a sizable fraction other creative professionals had the same problem, and couldn't beat it either, no matter how reasonable all the advice sounds.

"What do you do when you can't work?" my friends asked me. (Conversation probably not accurate, this is a very loose gist.)

And I replied that I usually browse random websites, or watch a short video.

"Well," they said, "if you know you can't work for a while, you should watch a movie or something."

"Unfortunately," I replied, "I have to do something whose time comes in short units, like browsing the Web or watching short videos, because I might become able to work again at any time, and I can't predict when—"

And then I stopped, because I'd just had a revelation.

I'd always thought of my workcycle as something *chaotic,* something *unpredictable.* I never used those words, but that was the way I *treated* it.

But here my friends seemed to be implying—what a strange thought—that *other* people could predict when they would become able to work again, and structure their time accordingly.

---

1. *http://lesswrong.com/lw/wb/chaotic_inversion/*
2. *http://lesswrong.com/lw/un/on_doing_the_impossible/*

And it occurred to me for the first time that I might have been committing that damned old chestnut the Mind Projection Fallacy[3], right out there in my ordinary everyday life instead of high abstraction.

Maybe it wasn't that my productivity was *unusually chaotic;* maybe I was just *unusually stupid* with respect to predicting it.

That's what inverted stupidity looks like—chaos. Something hard to handle, hard to grasp, hard to guess, something you can't do anything with. It's not just an idiom for high abstract things like Artificial Intelligence. It can apply in ordinary life too.

And the reason we don't think of the alternative explanation "I'm stupid", is *not*—I suspect—that we think so highly of ourselves. It's just that we don't think of ourselves at all. We just see a chaotic feature of the environment[4].

So now it's occurred to me that my productivity problem may not be chaos, but my own stupidity.

And that may or may not help anything. It certainly doesn't fix the problem right away. Saying "I'm ignorant" doesn't make you knowledgeable.

But it is, at least, a different path than saying "it's too chaotic".

3. *http://lesswrong.com/lw/oi/mind_projection_fallacy/*
4. *http://lesswrong.com/lw/oi/mind_projection_fallacy/*

Part III
How To Actually Change Your Mind

A sequence on the ultra-high-level penultimate technique of rationality: actually updating on evidence.

(Organized into eight subsequences.)

Politics is the Mind-Killer

A sequence on the various ways that politics damages our sanity — including, of course, making it harder to change our minds on political issues.

1. A Fable of Science and Politics[1]

In the time of the Roman Empire, civic life was divided between the Blue and Green factions. The Blues and the Greens murdered each other in single combats, in ambushes, in group battles, in riots. Procopius said of the warring factions: "So there grows up in them against their fellow men a hostility which has no cause, and at no time does it cease or disappear, for it gives place neither to the ties of marriage nor of relationship nor of friendship, and the case is the same even though those who differ with respect to these colors be brothers or any other kin." Edward Gibbon wrote: "The support of a faction became necessary to every candidate for civil or ecclesiastical honors."

Who were the Blues and the Greens? They were sports fans—the partisans of the blue and green chariot-racing teams.

Imagine a future society that flees into a vast underground network of caverns and seals the entrances. We shall not specify whether they flee disease, war, or radiation; we shall suppose the first Undergrounders manage to grow food, find water, recycle air, make light, and survive, and that their descendants thrive and eventually form cities. Of the world above, there are only legends written on scraps of paper; and one of these scraps of paper describes the *sky,* a vast open space of air above a great unbounded floor. The sky is cerulean in color, and contains strange floating objects like enormous tufts of white cotton. But the meaning of the word "cerulean" is controversial; some say that it refers to the color known as "blue", and others that it refers to the color known as "green".

In the early days of the underground society, the Blues and Greens contested with open violence; but today, truce prevails—a peace born of a growing sense of pointlessness. Cultural mores have changed; there is a large and prosperous middle class that has grown up with effective law enforcement and become unaccustomed to violence. The schools provide some

1. *http://lesswrong.com/lw/gt/a_fable_of_science_and_politics/*

sense of historical perspective; how long the battle between Blues and Greens continued, how many died, how little changed as a result. Minds have been laid open to the strange new philosophy that people are people, whether they be Blue or Green.

The conflict has not vanished. Society is still divided along Blue and Green lines, and there is a "Blue" and a "Green" position on almost every contemporary issue of political or cultural importance. The Blues advocate taxes on individual incomes, the Greens advocate taxes on merchant sales; the Blues advocate stricter marriage laws, while the Greens wish to make it easier to obtain divorces; the Blues take their support from the heart of city areas, while the more distant farmers and watersellers tend to be Green; the Blues believe that the Earth is a huge spherical rock at the center of the universe, the Greens that it is a huge flat rock circling some other object called a Sun. Not every Blue or every Green citizen takes the "Blue" or "Green" position on every issue, but it would be rare to find a city merchant who believed the sky was blue, and yet advocated an individual tax and freer marriage laws.

The Underground is still polarized; an uneasy peace. A few folk genuinely think that Blues and Greens should be friends, and it is now common for a Green to patronize a Blue shop, or for a Blue to visit a Green tavern. Yet from a truce originally born of exhaustion, there is a quietly growing spirit of tolerance, even friendship.

One day, the Underground is shaken by a minor earthquake. A sightseeing party of six is caught in the tremblor while looking at the ruins of ancient dwellings in the upper caverns. They feel the brief movement of the rock under their feet, and one of the tourists trips and scrapes her knee. The party decides to turn back, fearing further earthquakes. On their way back, one person catches a whiff of something strange in the air, a scent coming from a long-unused passageway. Ignoring the well-meant cautions of fellow travellers, the person borrows a

powered lantern and walks into the passageway. The stone corridor wends upward... and upward... and finally terminates in a hole carved out of the world, a place where all stone ends. Distance, endless distance, stretches away into forever; a gathering space to hold a thousand cities. Unimaginably far above, too bright to look at directly, a searing spark casts light over all visible space, the naked filament of some huge light bulb. In the air, hanging unsupported, are great incomprehensible tufts of white cotton. And the vast glowing ceiling above... the *color*... is...

Now history branches, depending on which member of the sightseeing party decided to follow the corridor to the surface.

Aditya the Blue stood under the blue forever, and slowly smiled. It was not a pleasant smile. There was hatred, and wounded pride; it recalled every argument she'd ever had with a Green, every rivalry, every contested promotion. *"You were right all along,"* the sky whispered down at her, *"and now you can prove it."* For a moment Aditya stood there, absorbing the message, glorying in it, and then she turned back to the stone corridor to tell the world. As Aditya walked, she curled her hand into a clenched fist. "The truce," she said, "is over."

Barron the Green stared incomprehendingly at the chaos of colors for long seconds. Understanding, when it came, drove a pile-driver punch into the pit of his stomach. Tears started from his eyes. Barron thought of the Massacre of Cathay, where a Blue army had massacred every citizen of a Green town, including children; he thought of the ancient Blue general, Annas Rell, who had declared Greens "a pit of disease; a pestilence to be cleansed"; he thought of the glints of hatred he'd seen in Blue eyes and something inside him cracked. *"How can you be on their side?"* Barron screamed at the sky, and then he began to weep; because he knew, standing under the malevolent blue glare, that the universe had always been a place of evil.

Charles the Blue considered the blue ceiling, taken aback. As a professor in a mixed college, Charles had carefully emphasized that Blue and Green viewpoints were equally valid and deserving of tolerance: The sky was a metaphysical construct, and cerulean a color that could be seen in more than one way. Briefly, Charles wondered whether a Green, standing in this place, might not see a green ceiling above; or if perhaps the ceiling would be green at this time tomorrow; but he couldn't stake the continued survival of civilization on that. This was merely a natural phenomenon of some kind, having nothing to do with moral philosophy or society... but one that might be readily misinterpreted, Charles feared. Charles sighed, and turned to go back into the corridor. Tomorrow he would come back alone and block off the passageway.

Daria, once Green, tried to breathe amid the ashes of her world. *I will not flinch,* Daria told herself, *I will not look away*. She had been Green all her life, and now she must be Blue. Her friends, her family, would turn from her. *Speak the truth, even if your voice trembles,* her father had told her; but her father was dead now, and her mother would never understand. Daria stared down the calm blue gaze of the sky, trying to accept it, and finally her breathing quietened. *I was wrong,* she said to herself mournfully; *it's not so complicated, after all.* She would find new friends, and perhaps her family would forgive her... or, she wondered with a tinge of hope, rise to this same test, standing underneath this same sky? "The sky is blue," Daria said experimentally, and nothing dire happened to her; but she couldn't bring herself to smile. Daria the Blue exhaled sadly, and went back into the world, wondering what she would say.

Eddin, a Green, looked up at the blue sky and began to laugh cynically. The course of his world's history came clear at last; even he couldn't believe they'd been such fools. "Stupid," Eddin said, "stupid, *stupid,* and all the

time it was right here." Hatred, murders, wars, and all along it was just a *thing* somewhere, that someone had written about like they'd write about any other thing. No poetry, no beauty, nothing that any sane person would ever care about, just one pointless thing that had been blown out of all proportion. Eddin leaned against the cave mouth wearily, trying to think of a way to prevent this information from blowing up the world, and wondering if they didn't all deserve it.

Ferris gasped involuntarily, frozen by sheer wonder and delight. Ferris's eyes darted hungrily about, fastening on each sight in turn before moving reluctantly to the next; the blue *sky,* the white *clouds,* the vast unknown *outside,* full of places and things (and people?) that no Undergrounder had ever seen. "Oh, so *that's* what color it is," Ferris said, and went exploring.

2. Politics is the Mind-Killer[1]

People go funny in the head when talking about politics. The evolutionary reasons for this are so obvious as to be worth belaboring: In the ancestral environment, politics was a matter of life and death. And sex, and wealth, and allies, and reputation... When, today, you get into an argument about whether "we" ought to raise the minimum wage, you're executing adaptations for an ancestral environment where being on the wrong side of the argument could get you killed. Being on the *right* side of the argument could let *you* kill your hated rival!

If you want to make a point about science, or rationality, then my advice is to not choose a domain from *contemporary* politics if you can possibly avoid it. If your point is inherently about politics, then talk about Louis XVI during the French Revolution. Politics is an important domain to which we should individually apply our rationality—but it's a terrible domain in which to *learn* rationality, or discuss rationality, unless all the discussants are already rational.

Politics is an extension of war by other means. Arguments are soldiers. Once you know which side you're on, you must support all arguments of that side, and attack all arguments that appear to favor the enemy side; otherwise it's like stabbing your soldiers in the back—providing aid and comfort to the enemy. People who would be level-headed about evenhandedly weighing all sides of an issue in their professional life as scientists, can suddenly turn into slogan-chanting zombies when there's a Blue or Green[2] position on an issue.

In Artificial Intelligence, and particularly in the domain of nonmonotonic reasoning, there's a standard problem: "All Quakers are pacifists. All Republicans are not pacifists. Nixon is a Quaker and a Republican. Is Nixon a pacifist?"

1. *http://lesswrong.com/lw/gw/politics_is_the_mindkiller/*
2. Page 143, 'A Fable of Science and Politics'.

What on Earth was the point of choosing this as an example? To rouse the political emotions of the readers and distract them from the main question? To make Republicans feel unwelcome in courses on Artificial Intelligence and discourage them from entering the field? (And no, before anyone asks, I am not a Republican. Or a Democrat.)

Why would anyone pick such a *distracting* example to illustrate nonmonotonic reasoning? Probably because the author just couldn't resist getting in a good, solid dig at those hated Greens[3]. It feels so *good* to get in a hearty punch, y'know, it's like trying to resist a chocolate cookie.

As with chocolate cookies, not everything that feels pleasurable is good for you. And it certainly isn't good for our hapless readers who have to read through all the angry comments your blog post inspired.

I'm not saying that I think *Overcoming Bias* should be apolitical, or even that we should adopt Wikipedia's ideal of the Neutral Point of View[4]. But try to resist getting in those good, solid digs if you can possibly avoid it. If your topic legitimately relates to attempts to ban evolution in school curricula, then go ahead and talk about it—but don't blame it explicitly on the whole Republican Party; some of your readers may be Republicans, and they may feel that the problem is a few rogues, not the entire party. As with Wikipedia's NPOV, it doesn't matter whether (you think) the Republican Party really *is* at fault. It's just better for the spiritual growth of the community to discuss the issue without invoking color politics[5].

(Now that I've been named as a co-moderator, I guess I'd better include a disclaimer: This article is my personal opinion, not a statement of official *Overcoming Bias* policy. This will always be the case unless explicitly specified otherwise.)

3. Page 143, 'A Fable of Science and Politics'.
4. *http://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view*
5. Page 143, 'A Fable of Science and Politics'.

3. Policy Debates Should Not Appear One-Sided[1]

Robin Hanson recently proposed stores where banned products could be sold[2]. There are a number of excellent arguments for such a policy—an inherent right of individual liberty, the career incentive of bureaucrats to prohibit *everything*, legislators being just as biased as individuals. But even so (I replied), *some* poor, honest, not overwhelmingly educated mother of 5 children is going to go into these stores and buy a "Dr. Snakeoil's Sulfuric Acid Drink" for her arthritis and die, leaving her orphans to weep on national television.

I was just making a simple factual observation. Why did some people think it was an argument in favor of regulation?

On questions of simple fact (for example, whether Earthly life arose by natural selection) there's a legitimate expectation that the argument should be a one-sided battle; the facts themselves are either one way or another, and the so-called "balance of evidence" should reflect this. Indeed, under the Bayesian definition of evidence, "strong evidence" is just that sort of evidence which we only expect to find on one side of an argument.

But there is no reason for complex actions with many consequences to exhibit this onesidedness property. Why do people seem to want their *policy* debates to be one-sided?

Politics is the mind-killer.[3] Arguments are soldiers. Once you know which side you're on, you must support all arguments of that side, and attack all arguments that appear to favor the enemy side; otherwise it's like stabbing your soldiers in the back. If you abide within that pattern, policy debates will also appear one-sided to you—the costs and drawbacks of your favored policy are enemy soldiers, to be attacked by any means necessary.

1. *http://lesswrong.com/lw/gz/policy_debates_should_not_appear_onesided/*
2. *http://www.overcomingbias.com/2007/03/paternalism_is_.html*
3. Page 148, 'Politics is the Mind-Killer'.

One should also be aware of a related failure pattern, thinking that the course of Deep Wisdom is to compromise with perfect evenness between whichever two policy positions receive the most airtime. A policy may legitimately have *lopsided* costs or benefits. If policy questions were not tilted one way or the other, we would be unable to make decisions about them. But there is also a human tendency to deny all costs of a favored policy, or deny all benefits of a disfavored policy; and people will therefore tend to think policy tradeoffs are tilted much further than they actually are.

If you allow shops that sell otherwise banned products, some poor, honest, poorly educated mother of 5 kids is going to buy something that kills her. This is a prediction about a factual consequence, and as a factual question it appears rather straightforward—a sane person should readily confess this to be true regardless of which stance they take on the policy issue. You may *also* think that making things illegal just makes them more expensive, that regulators will abuse their power, or that her individual freedom trumps your desire to meddle with her life. But, as a matter of simple fact, she's still going to die.

We live in an unfair universe. Like all primates, humans have strong negative reactions to perceived unfairness; thus we find this fact stressful. There are two popular methods of dealing with the resulting cognitive dissonance. First, one may change one's view of the facts—deny that the unfair events took place, or edit the history to make it appear fair. Second, one may change one's morality—deny that the events are unfair.

Some libertarians might say that if you go into a "banned products shop", passing clear warning labels that say "THINGS IN THIS STORE MAY KILL YOU", and buy something that kills you, then it's your own fault and you deserve it. If that were a moral truth, there would be *no downside* to having shops that sell banned products. It wouldn't just be a *net benefit*, it would be a *one-sided* tradeoff with no drawbacks.

Others argue that regulators can be trained to choose rationally and in harmony with consumer interests; if those were the facts of the matter then (in their moral view) there would be *no downside* to regulation.

Like it or not, there's a birth lottery for intelligence—though this is one of the cases where the universe's unfairness is so extreme that many people choose to deny the facts. The experimental evidence for a purely genetic component of 0.6-0.8 is overwhelming, but even if this were to be denied, you don't choose your parental upbringing or your early schools either.

I was raised to believe that denying reality is a *moral wrong*. If I were to engage in wishful optimism about how Sulfuric Acid Drink was likely to benefit me, I would be doing something that I was *warned* against and raised to regard as unacceptable. Some people are born into environments—we won't discuss their genes, because that part is too unfair—where the local witch doctor tells them that it is *right* to have faith and *wrong* to be skeptical. In all goodwill, they follow this advice and die. Unlike you, they weren't raised to believe that people are responsible for their individual choices to follow society's lead. Do you really think you're so smart that you would have been a proper scientific skeptic even if you'd been born in 500 C.E.? Yes, there is a birth lottery, no matter what you believe about genes.

Saying "People who buy dangerous products deserve to get hurt!" is not tough-minded. It is a way of refusing to live in an unfair universe. Real tough-mindedness is saying, "Yes, sulfuric acid is a horrible painful death, and no, that mother of 5 children didn't deserve it, but we're going to keep the shops open anyway because we did this cost-benefit calculation." Can you imagine a politician saying that? Neither can I. But insofar as economists have the power to influence policy, it might help if they could think it privately—maybe even say it in journal articles, suitably dressed up in polysyllabismic obfuscationalization so the media can't quote it.

I don't think that when someone makes a stupid choice and dies, this is a cause for celebration. I count it as a tragedy. It is not always helping people, to save them from the consequences of their own actions; but I draw a moral line at capital punishment. If you're dead, you can't learn from your mistakes.

Unfortunately the universe doesn't agree with me. We'll see which one of us is still standing when this is over.

ADDED: Two primary drivers of policy-one-sidedness are the affect heuristic[4] and the just-world fallacy[5].

4. Page 199, 'The Affect Heuristic'.
5. *http://en.wikipedia.org/wiki/Just-world_fallacy*

4. The Scales of Justice, the Notebook of Rationality[1]

Lady Justice[2] is widely depicted as carrying a scales. A scales has the property that whatever pulls one side down, pushes the other side up. This makes things very convenient and easy to track. It's also usually a gross distortion.

In human discourse there is a natural tendency to treat discussion as a form of combat, an extension of war, a sport; and in sports you only need to keep track of how many points have been scored by each team. There are only two sides[3], and every point scored against one side, is a point in favor of the other. Everyone in the audience keeps a mental running count of how many points each speaker scores against the other. At the end of the debate, the speaker who has scored more points is, obviously, the winner; so everything he says must be true, and everything the loser says must be wrong.

"The Affect Heuristic in Judgments of Risks and Benefits[4]" studied whether subjects mixed up their judgments of the possible benefits of a technology (e.g. nuclear power), and the possible risks of that technology, into a single overall good or bad feeling about the technology. Suppose that I first tell you that a particular kind of nuclear reactor generates less nuclear waste than competing reactor designs. But then I tell you that the reactor is more unstable than competing designs, with a greater danger of undergoing meltdown if a sufficiently large number of things go wrong simultaneously.

If the reactor is more likely to melt down, this seems like a 'point against' the reactor, or a 'point against' someone who argues for building the reactor. And if the reactor produces less

1. *http://lesswrong.com/lw/h1/the_scales_of_justice_the_notebook_of_rationality/*
2. *http://en.wikipedia.org/wiki/Lady_Justice*
3. Page 143, 'A Fable of Science and Politics'.
4. *http://www-abc.mpib-berlin.mpg.de/users/r20/finucane00_the_affect_heuristic.pdf*

waste, this is a 'point for' the reactor, or a 'point for' building it. So are these two facts opposed to each other? No. In the real world, no. These two facts may be cited by different sides of the same debate, but they are logically distinct; the facts don't know whose side they're on. The amount of waste produced by the reactor arises from physical properties of that reactor design. Other physical properties of the reactor make the nuclear reaction more unstable. Even if some of the same design properties are involved, you have to separately consider the probability of meltdown, and the expected annual waste generated. These are two different physical questions with two different factual answers.

But studies such as the above show that people tend to judge technologies—and many other problems—by an overall good or bad feeling. If you tell people a reactor design produces less waste, they rate its probability of meltdown as lower. This means getting the *wrong answer* to physical questions with definite factual answers, because you have mixed up logically distinct questions—treated facts like human soldiers on different sides of a war, thinking that any soldier on one side can be used to fight any soldier on the other side.

A scales is not wholly inappropriate for Lady Justice if she is investigating a strictly factual question of guilt or innocence. Either John Smith killed John Doe, or not. We are taught (by E. T. Jaynes) that all Bayesian evidence consists of probability flows *between* hypotheses; there is no such thing as evidence that "supports" or "contradicts" a single hypothesis, except insofar as other hypotheses do worse or better. So long as Lady Justice is investigating a *single*, strictly *factual* question with a *binary* answer space, a scales would be an appropriate tool. If Justitia must consider any more complex issue, she should relinquish her scales or relinquish her sword.

Not all arguments reduce to mere up or down. Lady Rationality carries a notebook, wherein she writes down all the facts that aren't on anyone's side.

5. Correspondence Bias[1]

> *The correspondence bias is the tendency to draw inferences about a person's unique and enduring dispositions from behaviors that can be entirely explained by the situations in which they occur.*
>
> —Gilbert and Malone[2]

We tend to see far too direct a correspondence between others' actions and personalities. When we see someone else kick a vending machine for no visible reason, we assume they are "an angry person". But when you yourself kick the vending machine, it's because the bus was late, the train was early, your report is overdue, and now the damned vending machine has eaten your lunch money for the second day in a row. *Surely,* you think to yourself, *anyone would kick the vending machine, in that situation.*

We attribute our own actions to our *situations,* seeing our behaviors as perfectly normal responses to experience. But when someone else kicks a vending machine, we don't see their past history trailing behind them in the air. We just see the kick, for no reason *we* know about, and we think this must be a naturally angry person—since they lashed out without any provocation.

Yet consider the prior probabilities. There are more late buses in the world, than mutants born with unnaturally high anger levels that cause them to sometimes spontaneously kick vending machines. Now the average human is, in fact, a mutant. If I recall correctly, an average individual has 2-10 somatically expressed mutations. But any *given* DNA location is very unlikely to be affected. Similarly, any given aspect of someone's

1. *http://lesswrong.com/lw/hz/correspondence_bias/*
2. *http://www.wjh.harvard.edu/~dtg/Gilbert%20&%20Malone%20(CORRESPONDENCE%20BIAS).pdf*

disposition is probably not very far from average. To suggest otherwise is to shoulder a burden of improbability.

Even when people are informed explicitly of situational causes, they don't seem to properly discount the observed behavior. When subjects are told that a pro-abortion or anti-abortion speaker was *randomly assigned* to give a speech on that position, subjects still think the speakers harbor leanings in the direction randomly assigned. (Jones and Harris 1967, "The attribution of attitudes.)

It seems quite intuitive to explain rain by water spirits; explain fire by a fire-stuff (phlogiston) escaping from burning matter; explain the soporific effect of a medication by saying that it contains a "dormitive potency". Reality usually involves more complicated mechanisms: an evaporation and condensation cycle underlying rain, oxidizing combustion underlying fire, chemical interactions with the nervous system for soporifics. But mechanisms sound more complicated than essences; they are harder to think of, less available. So when someone kicks a vending machine, we think they have an innate vending-machine-kicking-tendency.

Unless the "someone" who kicks the machine is us—in which case we're behaving perfectly normally, given our situations; surely anyone else would do the same. Indeed, we overestimate how likely others are to respond the same way we do—the "false consensus effect". Drinking students considerably overestimate the fraction of fellow students who drink, but nondrinkers considerably underestimate the fraction. The "fundamental attribution error" refers to our tendency to overattribute others' behaviors to their dispositions, while reversing this tendency for ourselves.

To understand why people act the way they do, we must first realize that everyone sees themselves as behaving normally. Don't ask what strange, mutant disposition they were born with, which directly corresponds to their surface behavior. Rather, ask what situations people see themselves as being

in. Yes, people do have dispositions—but there are not *enough* heritable quirks of disposition to directly account for all the surface behaviors you see.

Suppose I gave you a control with two buttons, a red button and a green button. The red button destroys the world, and the green button stops the red button from being pressed. Which button would you press? The green one. Anyone who gives a different answer is probably overcomplicating the question[3].

And yet people sometimes ask me why I want to save the world[4]. Like I must have had a traumatic childhood or something. Really, it seems like a pretty obvious decision... if you see the situation in those terms.

I may have non-average views which call for explanation—why do I believe such things, when most people don't?—but given those beliefs, my *reaction* doesn't seem to call forth an exceptional explanation. Perhaps I am a victim of false consensus; perhaps I overestimate how many people would press the green button if they saw the situation in those terms. But y'know, I'd still bet there'd be at least a *substantial minority*.

Most people see themselves as perfectly normal, from the inside. Even people you hate, people who do terrible things, are not exceptional mutants. No mutations are required, alas. When you understand this, you are ready to stop being surprised[5] by human events.

3. *http://intelligence.org/blog/2007/06/16/transhumanism-as-simplified-humanism/*
4. *http://intelligence.org/AIRisk.pdf*
5. *http://lesswrong.com/lw/hs/think_like_reality/*

6. Are Your Enemies Innately Evil?[1]

Followup to: Correspondence Bias[2]

As previously discussed[3], we see far too direct a correspondence between others' actions and their inherent dispositions. We see unusual dispositions that exactly match the unusual behavior, rather than asking after real situations or imagined situations that could explain the behavior. We hypothesize mutants.

When someone actually *offends* us—commits an action of which we (rightly or wrongly) disapprove—then, I observe, the correspondence bias redoubles. There seems to be a *very* strong tendency to blame evil deeds on the Enemy's mutant, evil disposition. Not as a moral point, but as a strict question of prior probability, we should ask what the Enemy might believe about their situation which would reduce the seeming bizarrity[4] of their behavior. This would allow us to hypothesize a less exceptional disposition, and thereby shoulder a lesser burden of improbability.

On September 11th, 2001, nineteen Muslim males hijacked four jet airliners in a deliberately suicidal effort to hurt the United States of America. Now why do you suppose they might have done that? Because they saw the USA as a beacon of freedom to the world, but were born with a mutant disposition that made them hate freedom?

Realistically, most people don't construct their life stories with themselves as the villains. Everyone is the hero of their own story. The Enemy's story, as seen by the Enemy, *is not going to make the Enemy look bad.* If you try to construe motivations that *would* make the Enemy look bad, you'll end up flat wrong about what actually goes on in the Enemy's mind.

1. *http://lesswrong.com/lw/io/are_your_enemies_innately_evil/*
2. Page 157, 'Correspondence Bias'.
3. Page 157, 'Correspondence Bias'.
4. *http://lesswrong.com/lw/hs/think_like_reality/*

But politics is the mind-killer.[5] Debate is war; arguments are soldiers. Once you know which side you're on, you must support all arguments of that side, and attack all arguments that appear to favor the opposing side; otherwise it's like stabbing your soldiers in the back.

If the Enemy did have an evil disposition, that would be an argument in favor of your side. And *any* argument that favors your side must be supported, no matter how silly—otherwise you're letting up the pressure somewhere on the battlefront. Everyone strives to outshine their neighbor in patriotic denunciation, and no one dares to contradict. Soon the Enemy has horns, bat wings, flaming breath, and fangs that drip corrosive venom. If you deny any aspect of this on merely factual grounds, you are arguing the Enemy's side; you are a traitor. Very few people will understand that you aren't defending the Enemy, just defending the truth.

If it took a mutant to do monstrous things, the history of the human species would look very different. Mutants would be rare.

Or maybe the fear is that understanding will lead to forgiveness. It's easier to shoot down evil mutants. It is a more inspiring battle cry to scream, "Die, vicious scum!" instead of "Die, people who could have been just like me but grew up in a different environment!" You might feel guilty killing people who *weren't* pure darkness.

This looks to me like the deep-seated yearning for a one-sided policy debate[6] in which the best policy has *no* drawbacks. If an army is crossing the border or a lunatic is coming at you with a knife, the policy alternatives are (a) defend yourself (b) lie down and die. If you defend yourself, you may have to kill. If you kill someone who could, in another world, have been your friend, that is a tragedy. And it *is* a tragedy. The other option, lying down and dying, is also a tragedy. Why must there be a

5. Page 148, 'Politics is the Mind-Killer'.
6. Page 150, 'Policy Debates Should Not Appear One-Sided'.

non-tragic option? Who says that the best policy available must have no downside? If someone has to die, it may as well be the initiator of force, to discourage future violence and thereby minimize the total sum of death.

If the Enemy has an average disposition, and is acting from beliefs about their situation that would make violence a typically human response, then that doesn't mean their beliefs are factually accurate. It doesn't mean they're justified. It means you'll have to shoot down someone who is the hero of their own story, and in their novel the protagonist will die on page 80. That is a tragedy, but it is better than the alternative tragedy. It is the choice that every police officer makes, every day, to keep our neat little worlds from dissolving into chaos.

When you accurately estimate the Enemy's psychology—when you know what is really in the Enemy's mind—that knowledge won't feel like landing a delicious punch on the opposing side[7]. It won't give you a warm feeling of righteous indignation. It won't make you feel good about yourself. If your estimate makes you feel unbearably sad, you may be seeing the world as it really is. More rarely, an accurate estimate may send shivers of serious horror down your spine, as when dealing with true psychopaths, or neurologically intact people with beliefs that have utterly destroyed their sanity (Scientologists or Jesus Camp[8]).

So let's come right out and say it—the 9/11 hijackers weren't evil mutants. They did not hate freedom. They, too, were the heroes of their own stories, and they died for what they believed was right—truth, justice, and the Islamic way. If the hijackers saw themselves that way, it doesn't mean their beliefs were true. If the hijackers saw themselves that way, it doesn't mean that we have to agree that what they did was justified. If the hijackers saw themselves that way, it doesn't mean that the passengers of United Flight 93 should have stood aside and let it

7. Page 148, 'Politics is the Mind-Killer'.
8. *http://www.youtube.com/watch?v=y_EKHK1C2IE*

happen. It does mean that in another world, if they had been raised in a different environment, those hijackers might have been police officers. And that is indeed a tragedy. Welcome to Earth.

7. The Robbers Cave Experiment[1]

Did you ever wonder, when you were a kid, whether your inane "summer camp" actually had some kind of elaborate hidden purpose—say, it was all a science experiment and the "camp counselors" were really researchers observing your behavior?

Me neither.

But we'd have been more paranoid if we'd read Intergroup Conflict and Cooperation: The Robbers Cave Experiment[2] by Sherif, Harvey, White, Hood, and Sherif (1954/1961). In this study, the experimental subjects—excuse me, "campers"—were 22 boys between 5th and 6th grade, selected from 22 different schools in Oklahoma City, of stable middle-class Protestant families, doing well in school, median IQ 112. They were as well-adjusted and as similar to each other as the researchers could manage.

The experiment, conducted in the bewildered aftermath of World War II, was meant to investigate the causes—and possible remedies—of intergroup conflict. How would they spark an intergroup conflict to investigate? Well, the 22 boys were divided into two groups of 11 campers, and—

—and that turned out to be quite sufficient.

The researchers' original plans called for the experiment to be conducted in three stages. In Stage 1, each group of campers would settle in, unaware of the other group's existence. Toward the end of Stage 1, the groups would gradually be made aware of each other. In Stage 2, a set of contests and prize competitions would set the two groups at odds.

They needn't have bothered with Stage 2. There was hostility almost from the moment each group became aware of the other group's existence: They were using *our* campground, *our* baseball diamond. On their first meeting, the two groups be-

1. *http://lesswrong.com/lw/lt/the_robbers_cave_experiment/*
2. *http://psychclassics.yorku.ca/Sherif/*

gan hurling insults. They named themselves the Rattlers and the Eagles (they hadn't needed names when they were the only group on the campground).

When the contests and prizes were announced, in accordance with pre-established experimental procedure, the intergroup rivalry rose to a fever pitch. Good sportsmanship in the contests was evident for the first two days but rapidly disintegrated.

The Eagles stole the Rattlers' flag and burned it. Rattlers raided the Eagles' cabin and stole the blue jeans of the group leader, which they painted orange and carried as a flag the next day, inscribed with the legend "The Last of the Eagles". The Eagles launched a retaliatory raid on the Rattlers, turning over beds, scattering dirt. Then they returned to their cabin where they entrenched and prepared weapons (socks filled with rocks) in case of a return raid. After the Eagles won the last contest planned for Stage 2, the Rattlers raided their cabin and stole the prizes. This developed into a fistfight that the staff had to shut down for fear of injury. The Eagles, retelling the tale among themselves, turned the whole affair into a magnificent victory—they'd chased the Rattlers "over halfway back to their cabin" (they hadn't).

Each group developed a negative stereotype of Them and a contrasting positive stereotype of Us. The Rattlers swore heavily. The Eagles, after winning one game, concluded that the Eagles had won because of their prayers and the Rattlers had lost because they used cuss-words all the time. The Eagles decided to stop using cuss-words themselves. They also concluded that since the Rattlers swore all the time, it would be wiser not to talk to them. The Eagles developed an image of themselves as proper-and-moral; the Rattlers developed an image of themselves as rough-and-tough.

Group members held their noses when members of the other group passed.

In Stage 3, the researchers tried to reduce friction between the two groups.

Mere contact (being present without contesting) did not reduce friction between the two groups. Attending pleasant events together—for example, shooting off Fourth of July fireworks—did not reduce friction; instead it developed into a food fight.

Would you care to guess what *did* work?

(Spoiler space...)

The boys were informed that there might be a water shortage in the whole camp, due to mysterious trouble with the water system—possibly due to vandals. (The Outside Enemy, one of the oldest tricks in the book.)

The area between the camp and the reservoir would have to be inspected by four search details. (Initially, these search details were composed uniformly of members from each group.) All details would meet up at the water tank if nothing was found. As nothing was found, the groups met at the water tank and observed for themselves that no water was coming from the faucet. The two groups of boys discussed where the problem might lie, pounded the sides of the water tank, discovered a ladder to the top, verified that the water tank was full, and finally found the sack stuffed in the water faucet. All the boys gathered around the faucet to clear it. Suggestions from members of both groups were thrown at the problem and boys from both sides tried to implement them.

When the faucet was finally cleared, the Rattlers, who had canteens, did not object to the Eagles taking a first turn at the faucets (the Eagles didn't have canteens with them). No insults were hurled, not even the customary "Ladies first".

It wasn't the end of the rivalry. There was another food fight, with insults, the next morning. But a few more common tasks, requiring cooperation from both groups—e.g. restarting a stalled truck—did the job. At the end of the trip, the Rattlers

used $5 won in a bean-toss contest to buy malts for all the boys in both groups.

The Robbers Cave Experiment illustrates the psychology of hunter-gatherer bands, echoed through time[3], as perfectly as any experiment ever devised by social science.

Any resemblance to modern politics is just your imagination.

(Sometimes I think humanity's second-greatest need is a supervillain. Maybe I'll go into that line of work after I finish my current job.)

Sherif, M., Harvey, O. J., White, B. J., Hood, W. R., & Sherif, C. W. 1954/1961. *Study of positive and negative intergroup attitudes between experimentally produced groups: Robbers Cave study*.[4] University of Oklahoma.

3. *http://lesswrong.com/lw/l1/evolutionary_psychology/*
4. *http://psychclassics.yorku.ca/Sherif/*

8. Reversed Stupidity Is Not Intelligence[1]

> "...then our people on that time-line went to work with corrective action. Here."
>
> He wiped the screen and then began punching combinations. Page after page appeared, bearing accounts of people who had claimed to have seen the mysterious disks, and each report was more fantastic than the last.
>
> "The standard smother-out technique," Verkan Vall grinned. "I only heard a little talk about the 'flying saucers,' and all of that was in joke. In that order of culture, you can always discredit one true story by setting up ten others, palpably false, parallel to it."
>
> —H. Beam Piper, *Police Operation*

Piper had a point. Pers'nally, I don't believe there are any poorly hidden aliens infesting these parts. But my disbelief has nothing to do with the awful embarrassing irrationality of flying saucer cults—at least, I hope not.

You and I believe that flying saucer cults arose in the total absence of any flying saucers. Cults can arise around almost any idea[2], thanks to human silliness. This silliness operates *orthogonally* to alien intervention: We would expect to see flying saucer cults whether or not there were flying saucers. Even if there were poorly hidden aliens, it would not be any *less* likely for flying saucer cults to arise. p(cults|aliens) isn't less than p(cults|~aliens), unless you suppose that poorly hidden aliens would deliberately suppress flying saucer cults. By the Bayesian definition of evidence[3], the observation "flying

1. *http://lesswrong.com/lw/lw/reversed_stupidity_is_not_intelligence/*
2. Page 247, 'Every Cause Wants To Be A Cult'.
3. *http://yudkowsky.net/rational/bayes*

saucer cults exist" is not evidence *against* the existence of flying saucers. It's not much evidence one way or the other.

This is an application of the general principle that, as Robert Pirsig puts it, "The world's greatest fool may say the Sun is shining, but that doesn't make it dark out."

If you knew someone who was wrong 99.99% of the time on yes-or-no questions, you could obtain 99.99% accuracy just by reversing their answers. They would need to do all the work of obtaining good evidence entangled with reality, and processing that evidence coherently, just to *anticorrelate* that reliably. They would have to be superintelligent to be that stupid.

A car with a broken engine cannot drive backward at 200 mph, even if the engine is *really really broken*.

If stupidity does not reliably anticorrelate with truth, how much less should human evil anticorrelate with truth? The converse of the halo effect[4] is the horns effect: All perceived negative qualities correlate. If Stalin is evil, then everything he says should be false. You wouldn't want to agree with *Stalin,* would you?

Stalin also believed that 2 + 2 = 4. Yet if you defend any statement made by Stalin, even "2 + 2 = 4", people will see only that you are "agreeing with Stalin"; you must be on his side.

Corollaries of this principle:

- To argue against an idea honestly, you should argue against the best arguments of the strongest advocates. Arguing against weaker advocates proves *nothing,* because even the strongest idea will attract weak advocates. If you want to argue against transhumanism or the intelligence explosion, you have to directly challenge the arguments of Nick Bostrom or Eliezer Yudkowsky post-2003. The least convenient path[5] is the only valid one.

4. Page 212, 'The Halo Effect'.
5. *http://lesswrong.com/lw/2k/the_least_convenient_possible_world/*

- Exhibiting sad, pathetic lunatics, driven to madness by their apprehension of an Idea, is no evidence against that Idea. Many New Agers have been made crazier by their personal apprehension of quantum mechanics[6].
- Someone once said, "Not all conservatives are stupid, but most stupid people are conservatives." If you cannot place yourself in a state of mind where this statement, true or false, seems *completely irrelevant* as a critique of conservatism, you are not ready to think rationally about politics.
- Ad hominem[7] argument is not valid.
- You need to be able to argue against genocide without saying "Hitler wanted to exterminate the Jews." If Hitler *hadn't* advocated genocide, would it thereby become okay?
- In Hansonian terms: Your instinctive willingness to believe something will change along with your willingness to *affiliate* with people who are known for believing it—quite apart from whether the belief is actually *true*. Some people may be reluctant to believe that God does not exist, not because there is evidence that God *does* exist, but rather because they are reluctant to affiliate with Richard Dawkins or those darned "strident" atheists who go around publicly saying "God does not exist".
- If your current computer stops working, you can't conclude that everything about the current system is wrong and that you need a new system without an AMD processor, an ATI video card, a Maxtor hard drive, or case fans—even though your current system has all these things and it doesn't work. Maybe you just need a new power cord.

6. *http://lesswrong.com/lw/r5/the_quantum_physics_sequence/*
7. *http://plover.net/~bonds/adhominem.html*

- If a hundred inventors fail[8] to build flying machines using metal and wood and canvas, it doesn't imply that what you really need is a flying machine of bone and flesh. If a thousand projects fail to build Artificial Intelligence using electricity-based computing, this doesn't mean that electricity is the source of the problem. Until you understand the problem, hopeful reversals are exceedingly unlikely to hit the solution[9].

8. *http://lesswrong.com/lw/vs/selling_nonapples/*
9. *http://lesswrong.com/lw/l9/artificial_addition/*

9. Argument Screens Off Authority[1]

Black Belt Bayesian[2] (aka "steven") tries to explain the asymmetry between good arguments and good authority, but it doesn't seem to be resolving the comments on Reversed Stupidity Is Not Intelligence[3], so let me take my own stab at it:

Scenario 1: Barry is a famous geologist. Charles is a fourteen-year-old juvenile delinquent with a long arrest record and occasional psychotic episodes. Barry flatly asserts to Arthur some counterintuitive statement about rocks, and Arthur judges it 90% probable. Then Charles makes an equally counterintuitive flat assertion about rocks, and Arthur judges it 10% probable. Clearly, Arthur is taking the speaker's *authority* into account in deciding whether to believe the speaker's assertions.

Scenario 2: David makes a counterintuitive statement about physics and gives Arthur a detailed explanation of the arguments, including references. Ernie makes an equally counterintuitive statement, but gives an unconvincing argument involving several leaps of faith. Both David and Ernie assert that this is the best explanation they can possibly give (to anyone, not just Arthur). Arthur assigns 90% probability to David's statement after hearing his explanation, but assigns a 10% probability to Ernie's statement.

It might seem like these two scenarios are roughly symmetrical: both involve taking into account useful evidence, whether strong versus weak authority, or strong versus weak argument.

But now suppose that Arthur asks Barry and Charles to make full technical cases, with references; and that Barry and Charles present equally good cases, and Arthur looks up the references and they check out. Then Arthur asks David and Ernie for their credentials, and it turns out that David and Ernie

1. *http://lesswrong.com/lw/lx/argument_screens_off_authority/*
2. *http://www.acceleratingfuture.com/steven/?p=33*
3. Page 168, 'Reversed Stupidity Is Not Intelligence'.

have roughly the same credentials—maybe they're both clowns, maybe they're both physicists.

Assuming that Arthur is knowledgeable enough to understand all the technical arguments—otherwise they're just impressive noises—it seems that Arthur should view David as having a great advantage in plausibility over Ernie, while Barry has at best a minor advantage over Charles.

Indeed, if the technical arguments are good enough, Barry's advantage over Charles may not be worth tracking. A good technical argument is one that *eliminates* reliance on the personal authority of the speaker.

Similarly, if we really believe Ernie that the argument he gave is the best argument he *could* give, which includes all of the inferential steps that Ernie executed, and all of the support that Ernie took into account—citing any authorities that Ernie may have listened to himself—then we can pretty much ignore any information about Ernie's credentials. Ernie can be a physicist or a clown, it shouldn't matter. (Again, this assumes we have enough technical ability to process the argument. Otherwise, Ernie is simply uttering mystical syllables, and whether we "believe" these syllables depends a great deal on his authority.)

So it seems there's an asymmetry between argument and authority. If we know authority we are still interested in hearing the arguments; but if we know the arguments fully, we have very little left to learn from authority.

Clearly (says the novice) authority and argument are fundamentally different kinds of evidence[4], a difference unaccountable in the boringly clean methods of Bayesian probability theory[5]. For while the strength of the evidences—90% versus 10%—is just the same in both cases, they do not behave similarly when combined. How, oh how, will we account for this?

4. Page 18, 'What is Evidence?'.
5. *http://yudkowsky.net/rational/bayes*

Here's half a technical demonstration of how to represent this difference in probability theory. (The rest you can take on my personal authority, or look up in the references.)

If p(H|E1) = 90% and p(H|E2) = 9%, what is the probability p(H|E1,E2)? If learning E1 is true leads us to assign 90% probability to H, and learning E2 is true leads us to assign 9% probability to H, then what probability should we assign to H if we learn both E1 and E2? This is simply not something you can calculate in probability theory from the information given. No, the missing information is not the prior probability of H. E1 and E2 may not be independent of each other.

Suppose that H is "My sidewalk is slippery", E1 is "My sprinkler is running", and E2 is "It's night." The sidewalk is slippery starting from 1 minute after the sprinkler starts, until just after the sprinkler finishes, and the sprinkler runs for 10 minutes. So if we know the sprinkler is on, the probability is 90% that the sidewalk is slippery. The sprinkler is on during 10% of the nighttime, so if we know that it's night, the probability of the sidewalk being slippery is 9%. If we know that it's night and the sprinkler is on—that is, if we know both facts—the probability of the sidewalk being slippery is 90%.

We can represent this in a graphical model as follows:

Night -> Sprinkler -> Slippery

Whether or not it's Night *causes* the Sprinkler to be on or off, and whether the Sprinkler is on *causes* the Sidewalk to be slippery or unslippery.

The direction of the arrows is meaningful. If I wrote:

Night -> Sprinkler <- Slippery

This would mean that, if I *didn't* know anything about the Sprinkler, the probability of Nighttime and Slipperiness would be independent of each other. For example, suppose that I roll Die One and Die Two, and add up the showing numbers to get the Sum:

Die 1 -> Sum <- Die 2.

If you don't tell me the sum of the two numbers, and you tell me the first die showed 6, this doesn't tell me anything about the result of the second die, yet. But if you now also tell me the sum is 7, I know the second die showed 1.

Figuring out when various pieces of information are dependent or independent of each other, given various background knowledge, actually turns into a quite technical topic. The books to read are Judea Pearl's Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference[6] and Causality[7]. (If you only have time to read one book, read the first one.)

If you know how to read causal graphs, then you look at the dice-roll graph and immediately see:

p(die1,die2) = p(die1)*p(die2)
p(die1,die2|sum) ≠ p(die1|sum)*p(die2|sum)

If you look at the correct sidewalk diagram, you see facts like:

p(slippery|night) ≠ p(slippery)
p(slippery|sprinkler) ≠ p(slippery)
p(slippery|night, sprinkler) = p(slippery|sprinkler)

That is, the probability of the sidewalk being Slippery, given knowledge about the Sprinkler and the Night, is the same probability we would assign if we knew only about the Sprinkler. Knowledge of the Sprinkler has made knowledge of the Night irrelevant to inferences about Slipperiness.

This is known as *screening off,* and the criterion that lets us read such conditional independences off causal graphs is known as *D-separation*.

For the case of argument and authority, the causal diagram looks like this:

Truth -> Argument Goodness -> Expert Belief

6. *http://www.amazon.com/Probabilistic-Reasoning-Intelligent-Systems-Plausible/dp/1558604790/*
7. *http://www.amazon.com/Causality-Reasoning-Inference-Judea-Pearl/dp/0521773628/*

If something is true, then it therefore tends to have arguments in favor of it, and the experts therefore observe these evidences and change their opinions. (In theory!)

If we see that an expert believes something, we infer back to the existence of evidence-in-the-abstract (even though we don't know what that evidence is exactly), and from the existence of this abstract evidence, we infer back to the truth of the proposition.

But if we know the value of the Argument node, this D-separates the node "Truth" from the node "Expert Belief" by blocking all paths between them, according to certain technical criteria for "path blocking" that seem pretty obvious in this case. So even without checking the exact probability distribution, we can read off from the graph that:

p(truth|argument,expert) = p(truth|argument)

This does not represent a contradiction of ordinary probability theory. It's just a more compact way of expressing certain probabilistic facts. You could read the same equalities and inequalities off an unadorned probability distribution—but it would be harder to see it by eyeballing. Authority and argument don't need two different kinds of probability, any more than sprinklers are made out of ontologically different stuff than sunlight.

In practice you can never *completely* eliminate reliance on authority. Good authorities are more likely to know about any counterevidence that exists and should be taken into account; a lesser authority is less likely to know this, which makes their arguments less reliable. This is not a factor you can eliminate merely by hearing the evidence they *did* take into account.

It's also very hard to reduce arguments to *pure* math; and otherwise, judging the strength of an inferential step may rely on intuitions you can't duplicate without the same thirty years of experience.

There is an ineradicable legitimacy to assigning *slightly* higher probability to what E. T. Jaynes tells you about Bayesian

probability, than you assign to Eliezer Yudkowsky making the exact same statement. Fifty additional years of experience should not count for literally *zero* influence.

But this slight strength of authority is only *ceteris paribus*, and can easily be overwhelmed by stronger arguments. I have a minor erratum in one of Jaynes's books—because algebra trumps authority.

10. Hug the Query[1]

Continuation of: Argument Screens Off Authority[2]

In the art of rationality there is a discipline of *closeness-to-the-issue*—trying to observe evidence that is as near to the original question as possible, so that it screens off as many other arguments as possible.

The Wright Brothers say, "My plane will fly." If you look at their authority (bicycle mechanics who happen to be excellent amateur physicists) then you will compare their authority to, say, Lord Kelvin, and you will find that Lord Kelvin is the greater authority.

If you demand to see the Wright Brothers' calculations, and you can follow them, and you demand to see Lord Kelvin's calculations (he probably doesn't have any apart from his own incredulity), then authority becomes much less relevant.

If you actually *watch the plane fly,* the calculations themselves become moot for many purposes, and Kelvin's authority not even worth considering.

The more *directly* your arguments bear on a question, without intermediate inferences—the closer the observed nodes are to the queried node, in the Great Web of Causality—the more powerful the evidence. It's a theorem of these causal graphs that you can never get *more* information from distant nodes, than from strictly closer nodes that screen off[3] the distant ones.

Jerry Cleaver said: "What does you in is not failure to apply some high-level, intricate, complicated technique. It's overlooking the basics. Not keeping your eye on the ball."

Just as it is superior to argue physics than credentials, it is also superior to argue physics than rationality. Who was more rational, the Wright Brothers or Lord Kelvin? If we can check

1. *http://lesswrong.com/lw/ly/hug_the_query/*
2. Page 172, 'Argument Screens Off Authority'.
3. Page 172, 'Argument Screens Off Authority'.

their calculations, we don't have to care! The virtue of a rationalist cannot *directly* cause a plane to fly.

If you forget this principle, learning about more biases will hurt you[4], because it will distract you from more direct arguments. It's all too easy to argue that someone is exhibiting Bias #182 in your repertoire of fully generic accusations, but you can't *settle* a factual issue without closer evidence. If there are biased reasons to say the sun is shining, that doesn't make it dark out.[5]

Just as you can't always experiment today[6], you can't always check the calculations today. Sometimes you don't know enough background material, sometimes there's private information, sometimes there just isn't time. There's a sadly large number of times when it's worthwhile to judge the speaker's rationality. You should always do it with a hollow feeling in your heart, though, a sense that something's missing.

Whenever you can, dance as near to the original question as possible—press yourself up against it—get close enough to *hug the query!*

4. Page 333, 'Knowing About Biases Can Hurt People'.
5. Page 168, 'Reversed Stupidity Is Not Intelligence'.
6. *http://lesswrong.com/lw/io/is_molecular_nanotechnology_scientific/*

11. Rationality and the English Language[1]

Yesterday[2], someone said that my writing reminded them of George Orwell's Politics and the English Language[3]. I was honored. Especially since I'd already thought of today's topic.

If you *really* want an artist's perspective[4] on rationality, then read Orwell; he is mandatory reading for rationalists as well as authors. Orwell was not a scientist, but a writer; his tools were not numbers, but words; his adversary was not Nature, but human evil. If you wish to imprison people for years without trial, you must think of some other way to say it than "I'm going to imprison Mr. Jennings for years without trial." You must muddy the listener's thinking, prevent clear images from outraging conscience. You say, "Unreliable elements were subjected to an alternative justice process."

Orwell was the outraged opponent of totalitarianism and the muddy thinking in which evil cloaks itself—which is how Orwell's writings on language ended up as classic rationalist documents on a level with Feynman, Sagan, or Dawkins.

"Writers are told to avoid usage of the passive voice." A rationalist whose background comes *exclusively* from science, may fail to see the flaw in the previous sentence; but anyone who's done a little writing should see it right away. I wrote the sentence in the passive voice, without telling you *who* tells authors to avoid passive voice. Passive voice removes the actor, leaving only the acted-upon. "Unreliable elements were subjected to an alternative justice process"—subjected by *who*? What does an "alternative justice process" *do*? With enough static noun phrases, you can keep anything unpleasant from actually *happening*.

---

1. *http://lesswrong.com/lw/jc/rationality_and_the_english_language/*
2. *http://lesswrong.com/lw/jb/applause_lights/f1t*
3. *http://www.k-1.com/Orwell/index.cgi/work/essays/language.html*
4. *http://lesswrong.com/lw/ja/we_dont_really_want_your_participation/*

Journal articles are often written in passive voice. (Pardon me, *some scientists* write their journal articles in passive voice. It's not as if the articles are being written by no one, with no one to blame.) It sounds more authoritative to say "The subjects were administered Progenitorivox" than "I gave each college student a bottle of 20 Progenitorivox, and told them to take one every night until they were gone." If you remove the scientist from the description, that leaves only the all-important data. But in reality the scientist *is* there, and the subjects *are* college students, and the Progenitorivox wasn't "administered" but handed over with instructions. Passive voice obscures reality.

Judging from the comments I get on Overcoming Bias, someone will protest that using the passive voice in a journal article is hardly a sin—after all, if you *think* about it, you can realize the scientist is there. It doesn't seem like a logical flaw. And this is why rationalists need to read Orwell, not just Feynman or even Jaynes.

Nonfiction conveys *knowledge,* fiction conveys *experience.* Medical science can extrapolate what would happen to a human unprotected in a vacuum. Fiction can make you live through it.

Some rationalists will try to analyze a misleading phrase[5], try to see if there *might possibly* be anything meaningful to it, try to *construct* a logical interpretation. They will be charitable, give the author the benefit of the doubt. Authors, on the other hand, are trained *not* to give themselves the benefit of the doubt. Whatever the audience *thinks* you said *is* what you said, whether you meant to say it or not; you can't argue with the audience no matter how clever your justifications.

A writer knows that readers will *not* stop for a minute to think. A fictional experience is a continuous stream of first impressions. A writer-rationalist pays attention to the *experience* words create. If you are evaluating the public rationality of a statement, and you analyze the words deliberatively, rephras-

5. Page 128, 'Applause Lights'.

ing propositions, trying out different meanings, searching for nuggets of truthiness, then you're losing track of the first impression—what the audience *sees,* or rather *feels*.

A novelist would notice the screaming wrongness of "The subjects were administered Progenitorivox." What life is here for a reader to live? This sentence creates a distant feeling of authoritativeness, and that's *all*—the *only* experience is the feeling of being told something reliable. A novelist would see nouns too abstract to show what actually happened—the postdoc with the bottle in his hand, trying to look stern; the student listening with a nervous grin.

My point is not to say that journal articles should be written like novels, but that a rationalist should become consciously aware of the *experiences* which words create. A rationalist must understand the mind and how to operate it. That includes the stream of consciousness, the part of yourself that unfolds in language. A rationalist must become consciously aware of the actual, experiential impact[6] of phrases, beyond their mere propositional semantics.

Or to say it more bluntly: *Meaning does not excuse impact!*

I don't care what rational interpretation you can *construct* for an applause light[7] like "AI should be developed through democratic processes". That cannot excuse its irrational impact of signaling the audience to applaud, not to mention its cloudy question-begging vagueness.

Here is Orwell, railing against the *impact* of cliches, their effect on the experience of thinking:

> When one watches some tired hack on the platform mechanically repeating the familiar phrases—-
> *bestial, atrocities, iron heel, bloodstained tyranny, free peoples of the world, stand shoulder to shoulder*
> —one often has a curious feeling that one is not

6. Page 92, 'Semantic Stopsigns'.
7. Page 128, 'Applause Lights'.

> watching a live human being but some kind of dummy... A speaker who uses that kind of phraseology has gone some distance toward turning himself into a machine. The appropriate noises are coming out of his larynx, but his brain is not involved, as it would be if he were choosing his words for himself...
>
> What is above all needed is to let the meaning choose the word, and not the other way around. In prose, the worst thing one can do with words is surrender to them. When you think of a concrete object, you think wordlessly, and then, if you want to describe the thing you have been visualising you probably hunt about until you find the exact words that seem to fit it. When you think of something abstract you are more inclined to use words from the start, and unless you make a conscious effort to prevent it, the existing dialect will come rushing in and do the job for you, at the expense of blurring or even changing your meaning. Probably it is better to put off using words as long as possible and get one's meaning as clear as one can through pictures and sensations.

Peirce[8] might have written that last paragraph. More than one path can lead to the Way.

8. *http://books.google.com/books?id=NK2dLn48zWIC&pg=PA338&lpg=PA338&ots=vDJyvFZotS&sig=5Z1ZG8_yfTFIUpCIq_YS*

12. The Litany Against Gurus[1]

I am your hero!
I am your master!
Learn my arts,
Seek my way.

Learn as I learned,
Seek as I sought.

Envy me!
Aim at me!
Rival me!
Transcend me!

Look back,
Smile,
And then—
Eyes front!

I was never your city,
Just a stretch of your road.

1. *http://lesswrong.com/lw/m2/the_litany_against_gurus/*

13. Politics and Awful Art[1]

Followup to: Rationality and the English Language[2]

One of my less treasured memories is of a State of the Union address, or possibly a presidential inauguration, at which a Nobel Laureate got up and read, in a terribly solemn voice, some politically correct screed about what a wonderfully inclusive nation we all were—"The African-Americans, the Ethiopians, the Etruscans", or something like that. The "poem", if you can call it that, was absolutely awful. As far as my ears could tell, it had no redeeming artistic merit whatsoever.

Every now and then, yet another atheist is struck by the amazing idea that atheists should have hymns, just like religious people have hymns, and they take some existing religious song and turn out an atheistic version. And then this "atheistic hymn" is, almost without exception, absolutely awful. But the author can't see how dreadful the verse is as verse. They're too busy congratulating themselves on having said "Religion sure sucks, amen." Landing a punch on the Hated Enemy feels so good that they overlook the hymn's lack of any *other* merit. Verse of the same quality about something unpolitical, like mountain streams, would be seen as something a kindergartener's mother would post on her refrigerator.

In yesterday's Litany Against Gurus[3], there are only two lines that might be classifiable as "poetry", not just "verse". When I was composing the litany's end, the lines that first popped into my head were:

> I was not your destination
> Only a step on your path

1. *http://lesswrong.com/lw/m3/politics_and_awful_art/*
2. Page 180, 'Rationality and the English Language'.
3. Page 184, 'The Litany Against Gurus'.

Which didn't sound right at all. Substitute "pathway" for "road", so the syllable counts would match? But that sounded even worse. The prosody—the pattern of stressed syllables—was all wrong.

The real problem was the word des-ti-NA-tion—a huge awkward lump four syllables long. So get rid of it! "I was not your goal" was the first alternative that came to mind. Nicely short. But now that I was thinking about it, "goal" sounded very airy and abstract. Then the word "city" came into my mind—and it echoed.

"I was never your city" came to me, not by thinking about rationality, but by thinking about prosody. The constraints of art force us to toss out the first, old, tired phrasing that comes to mind; and in searching for a less obvious phrasing, often lead us to less obvious thoughts.

If I'd said, "Well, this is such a wonderful thought about rationality, that I don't have to worry about the prosodic problem", then I would have not received the benefit of being constrained.

The other poetic line began as "Laugh once, and never look back," which had problems as rationality, not just as prosody. "Laugh once" is the wrong kind of laughter; too derisive. "Never look back" is even less correct, because the memory of past mistakes can be useful years later. So... "Look back, ~~laugh once~~ smile, and then," um, "look forward"? Now if I'd been enthralled by the wonders of rationality, I would have said, "Ooh, 'look forward'! What a progressive sentiment!" and forgiven the extra syllable.

"Eyes front!" It was two syllables. It had the crisp click of a drill sergeant telling you to stop woolgathering, snap out of that daze, and get to work! Nothing like the soft cliche of "look forward, look upward, look to the future in a vaguely admiring sort of way..."

Eyes front! It's a better thought as rationality, which I would never have found, if I'd been so impressed with daring to

write about rationality, that I had forgiven myself the prosodic transgression of an extra syllable.

If you allow affirmation of My-Favorite-Idea to compensate for lack of rhythm in a song, lack of beauty in a painting, lack of poignancy in fiction, then your art will, inevitably, suck. When you do art about My-Favorite-Idea, you have to hold yourself to the same standard as if you were doing art about a butterfly.

There is powerful politicized art, just as there are great religious paintings. But merit in politicized art is more the exception than the rule. Most of it ends up as New Soviet Man Heroically Crushing Capitalist Snakes. It's an easy living. If anyone criticizes your art on grounds of general suckiness, they'll be executed for siding with the capitalist snakes.

Tolerance of awful art, just because it lands a delicious punch on the Enemy, or just because it affirms the Great Truth, is a dangerous sign: It indicates an affective death spiral[4] entering the supercritical phase[5] where you can no longer criticize any argument whose conclusion is the "right" one.

And then the next thing you know, you're composing dreadful hymns, or inserting giant[6] philosophical[7] lectures[8] into the climax of your fictional novel...

4. Page 225, 'Affective Death Spirals'.
5. Page 235, 'Uncritical Supercriticality'.
6. *http://www.amazon.com/Atlas-Shrugged-Ayn-Rand/dp/0452011876/ref=pd_bbs_sr_1?ie=UTF8&s=books&qid=1198121423&sr=8-1*
7. *http://www.amazon.com/Wild-David-Zindell/dp/0553762192/ref=sr_1_1?ie=UTF8&s=books&qid=1198121477&sr=8-1*
8. *http://www.amazon.com/Golden-Transcendence-Last-Masquerade-Age/dp/0765349086/ref=pd_bbs_sr_1?ie=UTF8&s=books&qid=1198121504&sr=8-1*

14. False Laughter[1]

Followup to: Politics and Awful Art[2]

There's this thing called "derisive laughter" or "mean-spirited laughter", which follows from seeing the Hated Enemy get a kick in the pants. It doesn't have to be an unexpected kick in the pants, or a kick followed up with a custard pie. It suffices that the Hated Enemy gets hurt. It's like humor, only without the humor.

If you know what your audience hates, it doesn't take much effort to get a laugh like that—which marks this as a subspecies of awful political art[3].

There are deliciously biting satires, yes; not all political art is bad art. But satire is a much more demanding art than just punching the Enemy in the nose. In fact, never mind satire—just an atom of ordinary genuine humor takes effort.

Imagine this political cartoon: A building labeled "science", and a standard Godzilla-ish monster labeled "Bush" stomping on the "science" building. Now there are people who will laugh at this—hur hur, scored a point off Bush, hur hur—but this political cartoon didn't take much effort to imagine. In fact, it was *the very first example* that popped into my mind when I thought "political cartoon about Bush and science". This degree of obviousness is a bad sign.

If I want to make a *funny* political cartoon, I have to put in some effort. Go beyond the cached thought[4]. Use my *creativity*. Depict Bush as a tentacle monster and Science as a Japanese schoolgirl.

There are many art forms that suffer from obviousness. But humor more than most, because humor relies on surprise—the ridiculous, the unexpected, the absurd.

1. *http://lesswrong.com/lw/m5/false_laughter/*
2. Page 185, 'Politics and Awful Art'.
3. Page 185, 'Politics and Awful Art'.
4. Page 297, 'Cached Thoughts'.

(Satire achieves surprise by saying, out loud, the thoughts you didn't dare think. Fake satires repeat thoughts you were already thinking.)

You might say that a predictable punchline is too high-entropy to be funny, by that same logic which says you should be enormously less surprised to find your thermostat reading 30 degrees than 29 degrees.

The general test against awful political art is to ask whether the art would seem worthwhile if it were not political. If someone writes a song about space travel, and the song is good enough that I would enjoy listening to it even if it were about butterflies, then and only then does it qualify to pick up bonus points for praising a Worthy Cause.

So one test for derisive laughter is to ask if the joke would still be funny, if it weren't the Hated Enemy getting the kick in the pants. Bill Gates once got hit by an unexpected pie in the face. Would it still have been funny (albeit less funny) if Linus Torvalds had gotten hit by the pie?

Of course I'm not suggesting that you sit around all day asking which jokes are "really" funny, or which jokes you're "allowed" to laugh at. As the saying goes, analyzing a joke is like dissecting a frog—it kills the frog and it's not much fun for you, either.

So why this blog post, then? Don't you and I already know which jokes are funny?

One application: If you find yourself in a group of people who tell consistently unfunny jokes about the Hated Enemy, it may be a good idea to head for the hills, before you start to laugh as well...

Another application: You and I should be allowed *not* to laugh at certain jokes—even jokes that target our own favorite causes—on the grounds that the joke is too predictable to be funny. We should be able to do this without being accused of being humorless, "unable to take a joke", or protecting sacred cows. If labeled-Godzilla-stomps-a-labeled-building isn't fun-

ny about "Bush" and "Science", then it also isn't funny about "libertarian economists" and "American national competitiveness", etc.

The most scathing accusation I ever heard against Objectivism[5] is that hardcore Objectivists have no sense of humor; but no one could prove this by showing an Objectivist a cartoon of Godzilla-"Rand" stomping on building-"humor" and demanding that he laugh.

Requiring someone to laugh in order to prove their noncultishness—well, like most kinds of obligatory laughter, it doesn't quite work. Laughter, of all things, has to come naturally. The most you can do is get fear and insecurity *out of its way*.

If an Objectivist, innocently browsing the Internet, came across a depiction of Ayn Rand as a Japanese schoolgirl lecturing a tentacle monster, and *still* didn't laugh, then *that* would be a problem. But they couldn't fix this problem by deliberately trying to laugh.

Obstacles to humor are a sign of dreadful things. But making humor obligatory, or constantly wondering whether you're laughing enough, just throws up another obstacle. In that way it's rather Zen. There are things you can accomplish by deliberately composing a joke, but very few things you can accomplish by deliberately believing a joke is funny.

5. Page 258, 'Guardians of Ayn Rand'.

15. Human Evil and Muddled Thinking[1]

Followup to: Rationality and the English Language[2]

George Orwell[3] saw the descent of the civilized world into totalitarianism, the conversion or corruption of one country after another; the boot stamping on a human face, forever, and remember that it is forever. You were born too late to remember[4] a time when the rise of totalitarianism seemed unstoppable, when one country after another fell to secret police and the thunderous knock at midnight, while the professors of free universities hailed the Soviet Union's purges as progress. It feels as alien to you as fiction[5]; it is hard for you to take seriously[6]. Because, in your branch of time, the Berlin Wall fell. And if Orwell's name is not carved into one of those stones, it should be.

Orwell saw the destiny of the human species, and he put forth a convulsive effort to wrench it off its path. Orwell's weapon was clear writing. Orwell knew that muddled language is muddled thinking; he knew that human evil and muddled thinking intertwine like conjugate strands of DNA:

> In our time, political speech and writing are largely the defence of the indefensible. Things like the continuance of British rule in India, the Russian purges and deportations, the dropping of the atom bombs on Japan, can indeed be defended, but only by arguments which are too brutal for most people to face, and which do not square with the professed aims of the political parties. Thus political language

1. *http://lesswrong.com/lw/jd/human_evil_and_muddled_thinking/*
2. Page 180, 'Rationality and the English Language'.
3. *http://www.k-1.com/Orwell/index.cgi/work/essays/language.html*
4. Page 118, 'Making History Available'.
5. Page 118, 'Making History Available'.
6. Page 118, 'Making History Available'.

> has to consist largely of euphemism, question-begging and sheer cloudy vagueness. Defenceless villages are bombarded from the air, the inhabitants driven out into the countryside, the cattle machine-gunned, the huts set on fire with incendiary bullets: this is called *pacification*...

Orwell was clear on the goal of his clarity:

> If you simplify your English, you are freed from the worst follies of orthodoxy. You cannot speak any of the necessary dialects, and when you make a stupid remark its stupidity will be obvious, even to yourself.

To make our stupidity obvious, even to ourselves—this is the heart of Overcoming Bias.

Evil sneaks, hidden, through the unlit shadows of the mind. We look back with the clarity of history, and weep to remember[7] the planned famines of Stalin and Mao, which killed tens of millions[8]. We call this evil, because it was done by deliberate human intent to inflict pain and death upon innocent human beings. We call this evil, because of the revulsion that we feel against it, looking back with the clarity of history. For perpetrators of evil to avoid its natural opposition, the revulsion must remain latent. Clarity must be avoided at any cost. Even as humans of clear sight tend to oppose the evil that they see; so too does human evil, wherever it exists, set out to muddle thinking.

1984 sets this forth starkly: Orwell's ultimate villains are cutters and airbrushers of photographs (based on historical cutting and airbrushing in the Soviet Union). At the peak of all darkness in the Ministry of Love, O'Brien tortures Winston to admit that two plus two equals five:

7. Page 118, 'Making History Available'.
8. *http://lesswrong.com/lw/hw/scope_insensitivity/*

> 'Do you remember,' he went on, 'writing in your diary, "Freedom is the freedom to say that two plus two make four"?'
>
> 'Yes,' said Winston.
>
> O'Brien held up his left hand, its back towards Winston, with the thumb hidden and the four fingers extended.
>
> 'How many fingers am I holding up, Winston?'
>
> 'Four.'
>
> 'And if the party says that it is not four but five —then how many?'
>
> 'Four.'
>
> The word ended in a gasp of pain. The needle of the dial had shot up to fifty-five. The sweat had sprung out all over Winston's body. The air tore into his lungs and issued again in deep groans which even by clenching his teeth he could not stop. O'Brien watched him, the four fingers still extended. He drew back the lever. This time the pain was only slightly eased.

I am continually aghast at apparently intelligent folks—such as Robin's colleague Tyler Cowen[9]—who don't think that overcoming bias is important. This is your *mind* we're talking about. Your human intelligence[10]. It separates you from an ape. It built this world. You don't think how the mind works

9. *http://www.marginalrevolution.com/marginalrevolution/2007/08/how-important-i.html*
10. *http://intelligence.org/blog/2007/07/10/the-power-of-intelligence/*

is important? You don't think the mind's systematic malfunctions are important? Do you think the Inquisition would have tortured witches, if all were ideal Bayesians?

Tyler Cowen apparently feels that overcoming bias is just as biased as bias: "I view Robin's blog as exemplifying bias, and indeed showing that bias can be very useful." I *hope* this is only the result of thinking too abstractly while trying to sound clever. Does Tyler seriously think that scope insensitivity to the value of human life[11] is on the same level with trying to create plans that will *really* save as many lives as possible?

Orwell[12] was forced to fight a similar attitude—that to admit to any distinction is youthful naïveté:

> Stuart Chase and others have come near to claiming that all abstract words are meaningless, and have used this as a pretext for advocating a kind of political quietism. Since you don't know what Fascism is, how can you struggle against Fascism?

Maybe overcoming bias doesn't look quite exciting enough, if it's framed as a struggle against mere accidental mistakes. Maybe it's harder to get excited if there isn't some clear evil to oppose. So let us be absolutely clear that where there is human evil in the world, where there is cruelty and torture and deliberate murder, there are biases enshrouding it. Where people of clear sight oppose these biases, the concealed evil fights back. The truth *does* have enemies. If Overcoming Bias were a newsletter in the old Soviet Union, every poster and commenter of this blog would have been shipped off to labor camps.

In all human history, every great leap forward has been driven by a new clarity of thought. Except for a few natural catastrophes, every great woe has been driven by a stupidity.

---

11. *http://lesswrong.com/lw/hw/scope_insensitivity/*
12. *http://www.k-1.com/Orwell/index.cgi/work/essays/language.html*

Our last enemy is ourselves; and this is a war, and we are soldiers.

Death Spirals and the Cult Attractor

A subsequence of How to Actually Change Your Mind *on two of the huger obstacles, the affective death spiral and the cultishness attractor.*

Affective death spirals are positive feedback loop caused by the halo effect: Positive characteristics perceptually correlate, so the more nice things we say about X, the more additional nice things we're likely to believe about X.

Cultishness is an empirical attractor in human groups, roughly an affective death spiral, plus peer pressure and outcasting behavior, plus (quite often) defensiveness around something believed to have been perfected.

1. The Affect Heuristic[1]

The *affect heuristic* is when subjective impressions of goodness/badness act as a heuristic—a source of fast, perceptual judgments. Pleasant and unpleasant feelings are central to human reasoning, and the affect heuristic comes with lovely biases—some of my favorites.

Let's start with one of the relatively less crazy biases. You're about to move to a new city, and you have to ship an antique grandfather clock. In the first case, the grandfather clock was a gift from your grandparents on your 5th birthday. In the second case, the clock was a gift from a remote relative and you have no special feelings for it. How much would you pay for an insurance policy that paid out $100 if the clock were lost in shipping? According to Hsee and Kunreuther (2000), subjects stated willingness to pay more than twice as much in the first condition. This may sound rational—why not pay more to protect the more valuable object?—until you realize that the insurance doesn't *protect* the clock, it just pays if the clock is lost, and pays exactly the same amount for either clock. (And yes, it was stated that the insurance was with an outside company, so it gives no special motive to the movers.)

All right, but that doesn't *sound* too insane. Maybe you could get away with claiming the subjects were insuring affective outcomes, not financial outcomes—purchase of consolation.

Then how about this? Yamagishi (1997) showed that subjects judged a disease as more dangerous when it was described as killing 1,286 people out of every 10,000, versus a disease that was 24.14% likely to be fatal. Apparently the mental image of a thousand dead bodies is much more alarming, compared to a single person who's more likely to survive than not.

But wait, it gets worse.

1. *http://lesswrong.com/lw/lg/the_affect_heuristic/*

Suppose an airport must decide whether to spend money to purchase some new equipment, while critics argue that the money should be spent on other aspects of airport safety. Slovic et. al. (2002) presented two groups of subjects with the arguments for and against purchasing the equipment, with a response scale ranging from 0 (would not support at all) to 20 (very strong support). One group saw the measure described as saving 150 lives. The other group saw the measure described as saving 98% of 150 lives. The hypothesis motivating the experiment was that saving 150 lives sounds vaguely good—is that a lot? a little?—while saving 98% of something is clearly very good because 98% is so close to the upper bound of the percentage scale. Lo and behold, saving 150 lives had mean support of 10.4, while saving 98% of 150 lives had mean support of 13.6.

Or consider the report of Denes-Raj and Epstein (1994): Subjects offered an opportunity to win $1 each time they randomly drew a red jelly bean from a bowl, often preferred to draw from a bowl with more red beans and a smaller proportion of red beans. E.g., 7 in 100 was preferred to 1 in 10.

According to Denes-Raj and Epstein, these subjects reported afterward that even though they knew the probabilities were against them, they felt they had a better chance when there were more red beans. This may sound crazy to you, oh Statistically Sophisticated Reader, but if you think more carefully you'll realize that it makes perfect sense. A 7% probability versus 10% probability may be bad news, but it's more than made up for by the increased number of red beans. It's a worse probability, yes, but you're still more likely to *win,* you see. You should meditate upon this thought until you attain enlightenment as to how the rest of the planet thinks about probability.

Finucane et. al. (2000) tested the theory that people would conflate their judgments about particular good/bad aspects of something into an overall good or bad feeling about that thing. For example, information about a possible risk, or possible benefit, of nuclear power plants. Logically, information about risk

doesn't have to bear any relation to information about benefits. If it's a physical fact about a reactor design that it's passively safe (won't go supercritical even if the surrounding coolant systems and so on break down), this doesn't imply that the reactor will necessarily generate less waste, or produce electricity at a lower cost, etcetera. All these things would be good, but they are not the same good thing. Nonetheless, Finucane et. al. found that for nuclear reactors, natural gas, and food preservatives, presenting information about high benefits made people perceive lower risks; presenting information about higher risks made people perceive lower benefits; and so on across the quadrants.

Finucane et. al. also found that time pressure greatly *increased* the inverse relationship between perceived risk and perceived benefit, consistent with the general finding that time pressure, poor information, or distraction all increase the dominance of perceptual heuristics over analytic deliberation.

Ganzach (2001) found the same effect in the realm of finance. According to ordinary economic theory, return and risk should correlate *positively*—or to put it another way, people pay a premium price for safe investments, which lowers the return; stocks deliver higher returns than bonds, but have correspondingly greater risk. When judging *familiar* stocks, analysts' judgments of risks and returns were positively correlated, as conventionally predicted. But when judging *unfamiliar* stocks, analysts tended to judge the stocks as if they were generally good or generally bad—low risk and high returns, or high risk and low returns.

For further reading I recommend the fine summary chapter in Slovic et. al. 2002: "Rational Actors or Rational Fools: Implications of the Affect Heuristic for Behavioral Economics.[2]"

Denes-Raj, V., & Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their

2. *http://heuristics.behaviouralfinance.net/affect/Slov02.pdf*

better judgment. *Journal of Personality and Social Psychology,* 66, 819-829.

Finucane, M. L., Alhakami, A., Slovic, P., & Johnson, S. M. (2000). The affect heuristic in judgments of risks and benefits.[3] *Journal of Behavioral Decision Making,* 13, 1-17.

Ganzach, Y. (2001). Judging risk and return of financial assets. *Organizational Behavior and Human Decision Processes,* 83, 353-370.

Hsee, C. K. & Kunreuther, H. (2000). The affection effect in insurance decisions.[4] *Journal of Risk and Uncertainty*, 20, 141-159.

Slovic, P., Finucane, M., Peters, E. and MacGregor, D. 2002. Rational Actors or Rational Fools: Implications of the Affect Heuristic for Behavioral Economics.[5] *Journal of Socio-Economics,* 31: 329–342.

Yamagishi, K. (1997). When a 12.86% mortality is more dangerous than 24.14%: Implications for risk communication. *Applied Cognitive Psychology,* 11, 495-506.

3. *http://www-abc.mpib-berlin.mpg.de/users/r20/finucane00_the_affect_heuristic.pdf*
4. *http://faculty.chicagogsb.edu/christopher.hsee/vita/Papers/AffectionEffectInInsurance.pdf*
5. *http://heuristics.behaviouralfinance.net/affect/Slov02.pdf*

2. Evaluability (And Cheap Holiday Shopping)[1]

Followup to: The Affect Heuristic[2]

With the *expensive* part of the Hallowthankmas[3] season now approaching, a question must be looming large in our readers' minds:

> "Dear *Overcoming Bias,* are there biases I can exploit to be *seen* as generous without *actually* spending lots of money?"

I'm glad to report the answer is yes! According to Hsee (1998)—in a paper entitled "Less is better: When low-value options are valued more highly than high-value options"—if you buy someone a $45 scarf, you are more likely to be seen as generous than if you buy them a $55 coat.

This is a special case of a more general phenomenon. An earlier experiment, Hsee (1996), asked subjects how much they would be willing to pay for a second-hand music dictionary:

- Dictionary A, from 1993, with 10,000 entries, in like-new condition.
- Dictionary B, from 1993, with 20,000 entries, with a torn cover and otherwise in like-new condition.

The gotcha was that some subjects saw both dictionaries side-by-side, while other subjects only saw *one* dictionary...

Subjects who saw only *one* of these options were willing to pay an average of $24 for Dictionary A and an average of $20 for Dictionary B. Subjects who saw *both* options, side-by-side, were willing to pay $27 for Dictionary B and $19 for Dictionary A.

---

1. *http://lesswrong.com/lw/lh/evaluability_and_cheap_holiday_shopping/*
2. Page 199, 'The Affect Heuristic'.
3. *http://www.overcomingbias.com/2007/11/merry-hallowmas.html*

Of course, the number of entries in a dictionary is more important than whether it has a torn cover, at least if you ever plan on using it for anything. But if you're only presented with a single dictionary, and it has 20,000 entries, the number 20,000 doesn't mean very much. Is it a little? A lot? Who knows? It's *non-evaluable*. The torn cover, on the other hand—that stands out. That has a definite affective valence[4]: namely, bad.

Seen side-by-side, though, the number of entries goes from *non-evaluable* to *evaluable*, because there are two compatible quantities to be compared. And, once the number of entries becomes evaluable, that facet swamps the importance of the torn cover.

From Slovic et. al. (2002): Would you prefer:

1. A 29/36 chance to win $2
2. A 7/36 chance to win $9

While the average *prices* (equivalence values) placed on these options were $1.25 and $2.11 respectively, their mean attractiveness ratings were 13.2 and 7.5. Both the prices and the attractiveness rating were elicited in a context where subjects were told that two gambles would be randomly selected from those rated, and they would play the gamble with the higher price or higher attractiveness rating. (Subjects had a motive to rate gambles as more attractive, or price them higher, that they would actually prefer to play.)

The gamble worth more money seemed less attractive, a classic preference reversal. The researchers hypothesized that the dollar values were more compatible with the pricing task, but the probability of payoff was more compatible with attractiveness. So (the researchers thought) why not try to make the gamble's payoff more emotionally salient—more affectively evaluable—more attractive?

And how did they do this? By adding a very small loss to the gamble. The old gamble had a 7/36 chance of winning $9. The new gamble had a 7/36 chance of winning $9 and a 29/36

4. Page 199, 'The Affect Heuristic'.

chance of losing 5¢. In the old gamble, you implicitly evaluate the attractiveness of $9. The new gamble gets you to evaluate the attractiveness of winning $9 *versus* losing 5¢.

"The results," said Slovic. et. al., "exceeded our expectations." In a new experiment, the simple gamble with a 7/36 chance of winning $9 had a mean attractiveness rating of 9.4, while the complex gamble that included a 29/36 chance of losing 5¢ had a mean attractiveness rating of 14.9.

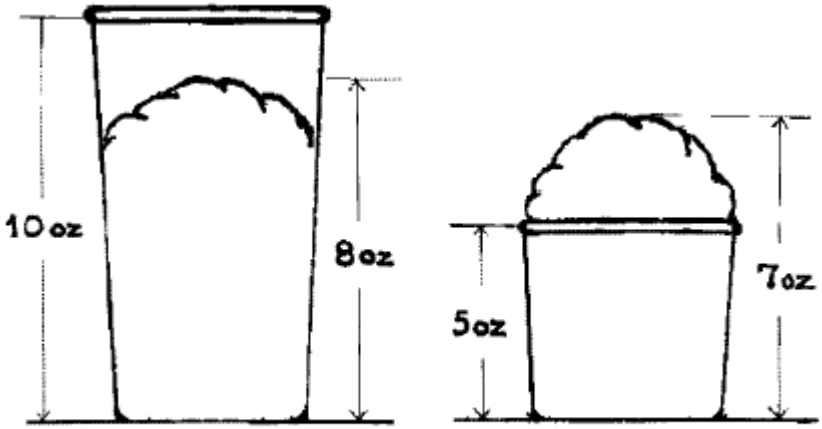
A follow-up experiment tested whether subjects preferred the old gamble to a certain gain of $2. Only 33% of students preferred the old gamble. Among another group asked to choose between a certain $2 and the new gamble (with the added possibility of a 5¢ loss), fully 60.8% preferred the gamble. After all, $9 isn't a very attractive amount of money, but $9/5¢ is an *amazingly* attractive win/loss ratio.

You can make a gamble more attractive by adding a strict loss! Isn't psychology fun? This is why no one who truly appreciates the wondrous[5] intricacy of human intelligence wants to design a human-like AI.

Of course, it only works if the subjects don't see the two gambles side-by-side.

Similarly, which of these two ice creams do you think subjects in Hsee (1998) preferred?

5. *http://lesswrong.com/lw/ks/the_wonder_of_evolution/*



Vendor H Vendor L

Naturally, the answer depends on whether the subjects saw a single ice cream, or the two side-by-side. Subjects who saw a single ice cream were willing to pay $1.66 to Vendor H and $2.26 to Vendor L. Subjects who saw both ice creams were willing to pay $1.85 to Vendor H and $1.56 to Vendor L.

What does this suggest for your holiday shopping? That if you spend $400 on a 16GB iPod Touch, your recipient sees the most expensive MP3 player. If you spend $400 on a Nintendo Wii, your recipient sees the least expensive game machine. Which is better value for the money? Ah, but that question only makes sense if you see the two side-by-side. *You'll* think about them side-by-side while you're shopping, but the recipient will only see what they get.

If you have a fixed amount of money to spend—and your goal is to display your friendship, rather than to actually *help* the recipient—you'll be better off deliberately not shopping for value. Decide how much money you want to spend on impressing the recipient, then find the most worthless object which

costs that amount. The cheaper the *class* of objects, the more expensive a *particular* object will appear, given that you spend a fixed amount. Which is more memorable, a $25 shirt or a $25 candle?

Gives a whole new meaning to the Japanese custom of buying $50 melons, doesn't it? You look at that and shake your head and say "What *is* it with the Japanese?". And yet they get to be perceived as incredibly generous, spendthrift even, while spending only $50. You could spend $200 on a fancy dinner and not appear as wealthy as you can by spending $50 on a melon. If only there was a custom of gifting $25 toothpicks or $10 dust specks; they could get away with spending even less.

PS: If you actually use this trick, I want to know what you bought.

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives.[6] *Organizational Behavior and Human Decision Processes,* 67, 242-257.

Hsee, C. K. (1998). Less is better: When low-value options are valued more highly than high-value options.[7] *Journal of Behavioral Decision Making,* 11, 107-121.

Slovic, P., Finucane, M., Peters, E. and MacGregor, D. (2002.) Rational Actors or Rational Fools: Implications of the Affect Heuristic for Behavioral Economics.[8] *Journal of Socio-Economics,* 31: 329–342.

6. *http://faculty.chicagogsb.edu/christopher.hsee/vita/Papers/EvaluabilityHypothesis.pdf*
7. *http://faculty.chicagogsb.edu/christopher.hsee/vita/Papers/LessIsBetter.pdf*
8. *http://heuristics.behaviouralfinance.net/affect/Slov02.pdf*

3. Unbounded Scales, Huge Jury Awards, & Futurism[1]

Followup to: Evaluability[2]

"Psychophysics", despite the name, is the respectable field that links physical effects to sensory effects. If you dump acoustic energy into air—make noise—then *how loud* does that sound to a person, as a function of acoustic energy? How much more acoustic energy do you have to pump into the air, before the noise sounds twice as loud to a human listener? It's not twice as much; more like eight times as much.

Acoustic energy and photons are straightforward to measure. When you want to find out how loud an acoustic stimulus *sounds,* how bright a light source *appears,* you usually ask the listener or watcher. This can be done using a bounded scale from "very quiet" to "very loud", or "very dim" to "very bright". You can also use an unbounded scale, whose zero is "not audible at all" or "not visible at all", but which increases from there without limit. When you use an unbounded scale, the observer is typically presented with a constant stimulus, the *modulus,* which is given a fixed rating. For example, a sound that is assigned a loudness of 10. Then the observer can indicate a sound twice as loud as the modulus by writing 20.

And this has proven to be a fairly reliable technique. But what happens if you give subjects an unbounded scale, but no modulus? 0 to infinity, with no reference point for a fixed value? Then they make up their own modulus, of course. The *ratios* between stimuli will continue to correlate reliably between subjects. Subject A says that sound X has a loudness of 10 and sound Y has a loudness of 15. If subject B says that sound X has a loudness of 100, then it's a good guess that subject B will assign loudness in the range of 150 to sound Y. But if you don't know what subject C is using as their modulus—their

1. *http://lesswrong.com/lw/li/unbounded_scales_huge_jury_awards_futurism/*
2. Page 203, 'Evaluability (And Cheap Holiday Shopping)'.

scaling factor—then there's no way to guess what subject C will say for sound X. It could be 1. It could be 1000.

For a subject rating a *single* sound, on an *unbounded* scale, *without* a fixed standard of comparison, nearly *all* the variance is due to the arbitrary choice of modulus, rather than the sound itself.

"Hm," you think to yourself, "this sounds an awful lot like juries deliberating on punitive damages. No wonder there's so much variance!" An interesting analogy, but how would you go about demonstrating it experimentally?

Kahneman et. al., 1998 and 1999, presented 867 jury-eligible subjects with descriptions of legal cases (e.g., a child whose clothes caught on fire) and asked them to either

1. Rate the outrageousness of the defendant's actions, on a bounded scale
2. Rate the degree to which the defendant should be punished, on a bounded scale, or
3. Assign a dollar value to punitive damages

And, lo and behold, while subjects correlated very well with each other in their outrage ratings and their punishment ratings, their punitive damages were all over the map. Yet subjects' *rank-ordering* of the punitive damages—their ordering from lowest award to highest award—correlated well across subjects.

If you asked how much of the variance in the "punishment" scale could be explained by the specific scenario—the particular legal case, as presented to multiple subjects—then the answer, even for the raw scores, was .49. For the *rank orders* of the dollar responses, the amount of variance predicted was .51. For the *raw dollar* amounts, the variance explained was .06!

Which is to say: if you knew the scenario presented—the aforementioned child whose clothes caught on fire—you could take a good guess at the punishment rating, and a good guess at the *rank-ordering* of the dollar award relative to other cases, but the dollar award itself would be completely unpredictable.

Taking the median of twelve randomly selected responses didn't help much either.

So a jury award for punitive damages isn't so much an economic valuation as an attitude expression—a psychophysical measure of outrage, expressed on an unbounded scale with no standard modulus.

I observe that many *futuristic predictions* are, likewise, best considered as attitude expressions. Take the question, "How long will it be until we have human-level AI?" The responses I've seen to this are all over the map. On one memorable occasion, a mainstream AI guy said to me, "Five hundred years." (!!)

Now the reason why time-to-AI is just *not very predictable,* is a long discussion in its own right. But it's not as if the guy who said "Five hundred years" was looking into the future to find out. And he can't have gotten the number using the standard bogus method with Moore's Law. So what did the number 500 *mean?*

As far as I can guess, it's as if I'd asked, "On a scale where zero is 'not difficult at all', how difficult does the AI problem *feel* to you?" If this were a bounded scale, every sane respondent would mark "extremely hard" at the right-hand end. Everything *feels* extremely hard when you don't know how to do it. But instead there's an unbounded scale with no standard modulus. So people just make up a number to represent "extremely difficult", which may come out as 50, 100, or even 500. Then they tack "years" on the end, and that's their futuristic prediction.

"How hard does the AI problem feel?" isn't the only substitutable question. Others respond as if I'd asked "How positive do you feel about AI?", only lower numbers mean more positive feelings, and then they also tack "years" on the end. But if these "time estimates" represent anything other than attitude expressions on an unbounded scale with no modulus, I have been unable to determine it.

Kahneman, D., Schkade, D. A., and Sunstein, C. 1998. Shared Outrage and Erratic Awards: The Psychology of Punitive Damages[3]. *Journal of Risk and Uncertainty* 16, 49-86.

Kahneman, D., Ritov, I. and Schkade, D. A. 1999. Economic Preferences or Attitude Expressions? An Analysis of Dollar Responses to Public Issues.[4] *Journal of Risk and Uncertainty,* 19: 203-235.

3. *http://ist-socrates.berkeley.edu/~maccoun/LP_KahnemanSchkadeSunstein1998.pdf*
4. *http://www.springerlink.com/content/u2322678545l4u6m/fulltext.pdf*

4. The Halo Effect[1]

The affect heuristic[2] is how an overall feeling of goodness or badness contributes to many other judgments, whether it's logical or not, whether you're aware of it or not. Subjects told about the benefits of nuclear power are likely to rate it as having fewer risks; stock analysts rating unfamiliar stocks judge them as generally good or generally bad—low risk and high returns, or high risk and low returns—in defiance of ordinary economic theory, which says that risk and return should correlate positively.

The halo effect is the manifestation of the affect heuristic[3] in social psychology. Robert Cialdini, in *Influence: Science and Practice*, summarizes:

> Research has shown that we automatically assign to good-looking individuals such favorable traits as talent, kindness, honesty, and intelligence (for a review of this evidence, see Eagly, Ashmore, Makhijani, & Longo, 1991). Furthermore, we make these judgments without being aware that physical attractiveness plays a role in the process. Some consequences of this unconscious assumption that "good-looking equals good" scare me. For example, a study of the 1974 Canadian federal elections found that attractive candidates received more than two and a half times as many votes as unattractive candidates (Efran & Patterson, 1976). Despite such evidence of favoritism toward handsome politicians, follow-up research demonstrated that voters did not realize their bias. In fact, 73 percent of Canadian voters surveyed denied in the strongest possible terms that their votes had been influenced by physical

1. *http://lesswrong.com/lw/lj/the_halo_effect/*
2. Page 199, 'The Affect Heuristic'.
3. Page 199, 'The Affect Heuristic'.

appearance; only 14 percent even allowed for the possibility of such influence (Efran & Patterson, 1976). Voters can deny the impact of attractiveness on electability all they want, but evidence has continued to confirm its troubling presence (Budesheim & DePaola, 1994).

A similar effect has been found in hiring situations. In one study, good grooming of applicants in a simulated employment interview accounted for more favorable hiring decisions than did job qualifications—this, even though the interviewers claimed that appearance played a small role in their choices (Mack & Rainey, 1990). The advantage given to attractive workers extends past hiring day to payday. Economists examining U.S. and Canadian samples have found that attractive individuals get paid an average of 12-14 percent more than their unattractive coworkers (Hammermesh & Biddle, 1994).

Equally unsettling research indicates that our judicial process is similarly susceptible to the influences of body dimensions and bone structure. It now appears that good-looking people are likely to receive highly favorable treatment in the legal system (see Castellow, Wuensch, & Moore, 1991; and Downs & Lyons, 1990, for reviews). For example, in a Pennsylvania study (Stewart, 1980), researchers rated the physical attractiveness of 74 separate male defendants at the start of their criminal trials. When, much later, the researchers checked court records for the results of these cases, they found that the handsome men had received significantly lighter sentences. In fact, attractive defendants were twice as likely to avoid jail as unattractive defendants. In

> another study—this one on the damages awarded in a staged negligence trial—a defendant who was better looking than his victim was assessed an average amount of $5,623; but when the victim was the more attractive of the two, the average compensation was $10,051. What's more, both male and female jurors exhibited the attractiveness-based favoritism (Kulka & Kessler, 1978).
>
> Other experiments have demonstrated that attractive people are more likely to obtain help when in need (Benson, Karabenic, & Lerner, 1976) and are more persuasive in changing the opinions of an audience (Chaiken, 1979)...

The influence of attractiveness on ratings of intelligence, honesty, or kindness is a clear example of bias—especially when you judge these other qualities based on fixed text—because we wouldn't expect judgments of honesty and attractiveness to conflate for any legitimate reason. On the other hand, how much of my perceived intelligence is due to my honesty? How much of my perceived honesty is due to my intelligence? Finding the truth, and saying the truth, are not as widely separated in nature as looking pretty and looking smart...

But these studies on the halo effect of attractiveness, should make us suspicious that there may be a similar halo effect for kindness, or intelligence. Let's say that you know someone who not only seems very intelligent, but also honest, altruistic, kindly, and serene. You should be suspicious that some of these perceived characteristics are influencing your perception of the others. Maybe the person is genuinely intelligent, honest, and altruistic, but not all that kindly[4] or serene[5]. You should be suspicious if the people you know seem to separate too cleanly into devils and angels.

4. *http://ozyandmillie.org/2003/03/24/ozy-and-millie-1134/*
5. *http://ozyandmillie.org/2006/11/16/ozy-and-millie-1770/*

And—I know you don't think *you* have to do it, but maybe *you* should—be just a little more skeptical of the more attractive political candidates.

Cialdini, R. B. 2001. *Influence: Science and Practice.* Boston, MA: Allyn and Bacon.

Cialdini's references:

Benson, P. L., Karabenic, S. A., & Lerner, R. M. (1976). Pretty pleases: The effects of physical attractiveness, race, and sex on receiving help. *Journal of Experimental Social Psychology, 12,* 409-415.

Budesheim, T. L., & DePaola, S. J. (1994). Beauty or the beast? The effects of appearance, personality, and issue information on evaluations of political candidates. *Personality and Social Psychology Bulletin, 20,* 339-348.

Castellow, W. A., Wuensch, K. L., & Moore, C. H. (1990). Effects of physical attractiveness of the plaintiff and defendant in sexual harassment judgments. *Journal of Social Behavior and Personality, 5,* 547-562.

Chaiken, S. ((1979). Communicator physical attractiveness and persuasion. *Journal of Personality and Social Psychology, 5,* 547-562.

Downs, A. C., & Lyons, P. M. (1990). Natural observations of the links between attractiveness and initial legal judgments. *Personality and Social Psychology Bulletin, 17,* 541-547.

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but...: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin, 110,* 109-128.

Efran, M. G., & Patterson, E. W. J. (1976). The politics of appearance. Unpublished manuscript, University of Toronto.

Hammermesh, D., & Biddle, J. E. (1994). Beauty and the labor market. *The American Economic Review, 84,* 1174-1194.

Kulka, R. A., & Kessler, J. R. (1978). Is justice really blind? The effect of litigant physical attractiveness on judicial judgment. *Journal of Applied Social Psychology, 4,* 336-381.

Mack, D., & Rainey, D. (1990). Female applicants' grooming and personnel selection. *Journal of Social Behavior and Personality, 5,* 399-407.

Stewart, J. E. II. (1980). Defendant's attractiveness as a factor in the outcome of trials. *Journal of Applied Social Psychology, 10,* 348-361.

5. Superhero Bias[1]

Followup to: The Halo Effect[2]

Suppose there's a heavily armed sociopath, a kidnapper with hostages, who has just rejected all requests for negotiation and announced his intent to start killing. In real life, the good guys don't usually kick down the door when the bad guy has hostages. But sometimes—*very* rarely, but sometimes—life imitates Hollywood to the extent of genuine good guys needing to smash through a door.

Imagine, in two widely separated realities, two heroes who charge into the room, first to confront the villain.

In one reality, the hero is strong enough to throw cars, can fire power blasts out of his nostrils, has X-ray hearing, and his skin doesn't just *deflect* bullets but annihilates them on contact. The villain has ensconced himself in an elementary school and taken over two hundred children hostage; their parents are waiting outside, weeping.

In another reality, the hero is a New York police officer, and the hostages are three prostitutes the villain collected off the street.

Consider this question very carefully: Who is the greater hero? And who is more likely to get their own comic book?

The halo effect[3] is that perceptions of all positive traits are correlated. Profiles rated higher on scales of attractiveness, are also rated higher on scales of talent, kindness, honesty, and intelligence.

And so comic-book characters who seem strong and invulnerable, both positive traits, also seem to possess more of the heroic traits of courage and heroism. And yet:

1. *http://lesswrong.com/lw/lk/superhero_bias/*
2. Page 212, 'The Halo Effect'.
3. Page 212, 'The Halo Effect'.

> "How tough can it be to act all brave and courageous when you're pretty much invulnerable?"
> —*Empowered,* Vol. 1

I can't remember if I read the following point somewhere, or hypothesized it myself: *Fame,* in particular, seems to combine additively with all other personality characteristics. Consider Gandhi. Was Gandhi the *most altruistic* person of the 20th century, or just the *most famous* altruist? Gandhi faced police with riot sticks and soldiers with guns. But Gandhi was a celebrity, and he was protected by his celebrity. What about the others in the march, the people who faced riot sticks and guns even though there wouldn't be international headlines if they were put in the hospital or gunned down?

What did Gandhi think of getting the headlines, the celebrity, the fame, the place in history, *becoming the archetype* for non-violent resistance, when he took less risk than any of the people marching with him? How did he feel when one of those anonymous heroes came up to him, eyes shining, and told Gandhi how wonderful he was? Did Gandhi ever visualize his world in those terms? I don't know; I'm not Gandhi.

This is not in any sense a criticism of Gandhi. The point of non-violent resistance is not to show off your courage. That can be done much more easily by going over Niagara Falls in a barrel. Gandhi couldn't help being somewhat-but-not-entirely protected by his celebrity. And Gandhi's actions did take courage—not as much courage as marching anonymously, but still a great deal of courage.

The bias I wish to point out is that Gandhi's fame score seems to get perceptually *added* to his justly accumulated altruism score. When you think about nonviolence, you think of Gandhi—not an anonymous protestor in one of Gandhi's marches who faced down riot clubs and guns, and got beaten, and had to be taken to the hospital, and walked with a limp for the rest of her life, *and no one ever remembered her name.*

Similarly, which is greater—to risk your life to save two hundred children, or to risk your life to save three adults?

The answer depends on what one means by *greater*. If you ever have to *choose* between saving three adults and saving two hundred children, then choose the latter. "Whoever saves a single life, it is as if he had saved the whole world[4]" may be a fine applause light[5], but it's terrible moral advice if you've got to pick one or the other. So if you mean "greater" in the sense of "Which is more important?" or "Which is the preferred outcome?" or "Which should I choose if I have to do one or the other?" then it is greater to save two hundred than three.

But if you ask about greatness in the sense of revealed virtue, then someone who would risk their life to save only three lives, reveals more courage than someone who would risk their life to save two hundred but not three.

This doesn't mean that you can deliberately choose to risk your life to save three adults, and let the two hundred schoolchildren go hang, because you want to reveal more virtue. Someone who risks their life *because they want to be virtuous* has revealed far less virtue than someone who risks their life *because they want to save others*. Someone who chooses to save three lives rather than two hundred lives, because they think it reveals greater virtue, is so selfishly fascinated with their own "greatness" as to have committed the moral equivalent of manslaughter.

It's one of those *wu wei* scenarios: You cannot reveal virtue by trying to reveal virtue. Given a choice between a safe method to save the world which involves no personal sacrifice or discomfort, and a method that risks your life and requires you to endure great privation, you cannot become a hero by deliberately choosing the second path. There is nothing heroic about wanting to be a hero. It would be a lost purpose[6].

4. *http://lesswrong.com/lw/hx/one_life_against_the_world/*
5. Page 128, 'Applause Lights'.
6. *http://lesswrong.com/lw/le/lost_purposes/*

Truly virtuous people who are genuinely trying to save lives, rather than trying to reveal virtue, will constantly seek to save more lives with less effort, which means that less of their virtue will be revealed. It may be confusing, but it's not contradictory.

But we cannot always choose to be invulnerable to bullets. After we've done our best to reduce risk and increase scope, any *remaining* heroism is well and truly revealed.

The police officer who puts their life on the line with no superpowers, no X-Ray vision, no super-strength, no ability to fly, and above all no invulnerability to bullets, reveals far greater virtue than Superman—who is only a *mere superhero*.

6. Mere Messiahs[1]

Followup to: Superhero Bias[2]

Yesterday I discussed how the halo effect[3], which causes people to see all positive characteristics as correlated—for example, more attractive individuals are also perceived as more kindly, honest, and intelligent—causes us to admire heroes more if they're super-strong and immune to bullets. Even though, logically, it takes much more courage to be a hero if you're *not* immune to bullets. Furthermore, it reveals more virtue to act courageously to save one life than to save the world. (Although if you have to do one or the other, of course you should save the world[4].)

"The police officer who puts their life on the line with no superpowers", I said, "reveals far greater virtue than Superman, who is a *mere superhero*."

But let's be more specific.

John Perry[5] was a New York City police officer who also happened to be an Extropian and transhumanist, which is how I come to know his name. John Perry was due to retire shortly and start his own law practice, when word came that a plane had slammed into the World Trade Center. He died when the north tower fell. I didn't know John Perry personally, so I cannot attest to this of direct knowledge; but very few Extropians believe in God, and I expect that Perry was likewise an atheist.

Which is to say that Perry knew he was risking his very existence, every week on the job. And it's not, like most people in history, that he knew he had only a choice of how to die, and chose to make it matter—because Perry was a transhumanist; he had genuine hope. And Perry went out there and put his

1. *http://lesswrong.com/lw/ll/mere_messiahs/*
2. Page 217, 'Superhero Bias'.
3. Page 212, 'The Halo Effect'.
4. *http://lesswrong.com/lw/hx/one_life_against_the_world/*
5. *http://www.nleomf.org/911heroes/Perry.html*

life on the line anyway. Not because he expected any divine reward. Not because he expected to experience anything at all, if he died. But because there were other people in danger, and they didn't have immortal souls either, and his hope of life was worth no more than theirs.

I did not know John Perry. I do not know if he saw the world this way. But the fact that an atheist and a transhumanist can still be a police officer, can still run into the lobby of a burning building, says more about the human spirit than all the martyrs who ever hoped of heaven.

So that is one specific police officer...

...and now for the superhero.

As the Christians tell the story, Jesus Christ could walk on water, calm storms, drive out demons with a word. It must have made for a comfortable life: Starvation a problem? Xerox some bread. Don't like a tree? Curse it. Romans a problem? Sic your Dad on them. Eventually this charmed life ended, when Jesus voluntarily presented himself for crucifixion. Being nailed to a cross is not a comfortable way to die. But as the Christians tell the story, Jesus did this knowing he would come back to life three days later, and then go to Heaven. What was the threat that moved Jesus to face this temporary suffering followed by eternity in Heaven? Was it the life of a single person? Was it the corruption of the church of Judea, or the oppression of Rome? No: as the Christians tell the story, the eternal fate of every human went on the line before Jesus suffered himself to be temporarily nailed to a cross.

But I do not wish to condemn a man who is not truly so guilty. What if Jesus—no, let's pronounce his name correctly: Yeishu—what if Yeishu of Nazareth never walked on water, and *nonetheless* defied the church of Judea established by the powers of Rome?

Would that not deserve greater honor than that which adheres to Jesus Christ, who was only a mere messiah?

Alas, somehow it seems greater for a hero to have steel skin and godlike powers. Somehow it seems to reveal more virtue to die temporarily to save the whole world, than to die permanently confronting a corrupt church. It seems so *common,* as if many other people through history had done the same.

Comfortably ensconced two thousand years in the future, we can levy all sorts of criticisms at Yeishu, but Yeishu did what he believed to be right, confronted a church he believed to be corrupt, and died for it. Without benefit of hindsight[6], he could hardly be expected to predict the true impact of his life upon the world. Relative to most other prophets of his day, he was probably relatively more honest, relatively less violent, and relatively more courageous. If you strip away the unintended consequences, the worst that can be said of Yeishu is that others in history did better. (Epicurus, Buddha, and Marcus Aurelius all come to mind.) Yeishu died forever, and—from one perspective—he did it for the sake of honesty. Fifteen hundred years before science, religious honesty was not an oxymoron.

As Sam Harris said:

> "It is not enough that Jesus was a man who transformed himself to such a degree that the Sermon on the Mount could be his heart's confession. He also had to be the Son of God, born of a virgin, and destined to return to earth trailing clouds of glory. The effect of such dogma is to place the example of Jesus forever out of reach. His teaching ceases to become a set of empirical claims about the linkage between ethics and spiritual insight and instead becomes a gratuitous, and rather gruesome, fairy tale. According to the dogma of Christianity, becoming just like Jesus is impossible. One can only enumerate one's sins, believe the unbelievable, and await the end of the world."

6. Page 71, 'Hindsight bias'.

I severely doubt that Yeishu ever spoke the Sermon on the Mount. Nonetheless, Yeishu deserves honor. He deserves more honor than the Christians would grant him.

But since Yeishu probably anticipated[7] his soul would survive, he doesn't deserve more honor than John Perry.

7. Page 43, 'Belief in Belief'.

7. Affective Death Spirals[1]

Followup to: The Affect Heuristic[2], The Halo Effect[3]

Many[4], many[5], many[6] are the flaws in human reasoning which lead us to overestimate how well our beloved theory explains the facts. The phlogiston theory of chemistry could explain just about anything, so long as it didn't have to predict it in advance. And the more phenomena you use your favored theory to explain, the truer your favored theory seems—has it not been confirmed by these many observations? As the theory seems truer, you will be more likely to question evidence that conflicts with it. As the favored theory seems more general, you will seek to use it in more explanations.

If you know anyone who believes that Belgium secretly controls the US banking system, or that they can use an invisible blue spirit force to detect available parking spaces, that's probably how they got started.

(Just keep an eye out, and you'll observe much that seems to confirm this theory...)

This positive feedback cycle of credulity and confirmation is indeed fearsome, and responsible for much error, both in science and in everyday life.

But it's nothing compared to the death spiral that begins with a charge of positive affect—a thought that *feels really good.*

A new political system that can save the world. A great leader, strong and noble and wise. An amazing tonic that can cure upset stomachs and cancer.

---

1. *http://lesswrong.com/lw/lm/affective_death_spirals/*
2. Page 199, 'The Affect Heuristic'.
3. Page 212, 'The Halo Effect'.
4. Page 71, 'Hindsight bias'.
5. Page 87, 'Fake Causality'.
6. Page 351, 'Rationalization'.

Heck, why not go for all three? A great cause needs a great leader. A great leader should be able to brew up a magical tonic or two.

The halo effect[7] is that any perceived positive characteristic (such as attractiveness or strength) increases perception of any other positive characteristic (such as intelligence or courage). Even when it makes no sense, or less than no sense[8].

Positive characteristics enhance perception of every other positive characteristic? That sounds a lot like how a fissioning uranium atom sends out neutrons that fission other uranium atoms.

Weak positive affect is subcritical; it doesn't spiral out of control. An attractive person seems more honest, which, perhaps, makes them seem more attractive; but the effective neutron multiplication factor is less than 1. Metaphorically speaking. The resonance confuses things a little, but then dies out.

With intense positive affect attached to the Great Thingy, the resonance touches everywhere. A believing Communist sees the wisdom of Marx in every hamburger bought at McDonalds; in every promotion they're denied that would have gone to them in a true worker's paradise; in every election that doesn't go to their taste, in every newspaper article "slanted in the wrong direction". Every time they use the Great Idea to interpret another event, the Great Idea is confirmed all the more. It feels better—positive reinforcement—and of course, when something feels good, that, alas, makes us *want* to believe it all the more.

When the Great Thingy feels good enough to make you *seek out* new opportunities to feel even better about the Great Thingy, applying it to interpret new events every day, the resonance of positive affect is like a chamber full of mousetraps loaded with ping-pong balls[9].

7. Page 212, 'The Halo Effect'.
8. Page 217, 'Superhero Bias'.
9. *http://www.youtube.com/watch?v=ORqc1x3_Evg&feature=related*

You could call it a "happy attractor", "overly positive feedback", a "praise locked loop", or "funpaper". Personally I prefer the term "affective death spiral".

Coming tomorrow: How to resist an affective death spiral. (Hint: It's not by refusing to ever admire anything again, nor by keeping the things you admire in safe little restricted magisteria.)

8. Resist the Happy Death Spiral[1]

Followup to: Affective Death Spirals[2]

Once upon a time, there was a man who was convinced that he possessed a Great Idea. Indeed, as the man thought upon the Great Idea more and more, he realized that it was not just *a* great idea, but *the most wonderful idea ever*. The Great Idea would unravel the mysteries of the universe, supersede the authority of the corrupt and error-ridden Establishment, confer nigh-magical powers upon its wielders, feed the hungry, heal the sick, make the whole world a better place, etc. etc. etc.

The man was Francis Bacon, his Great Idea was the scientific method, and he was the only crackpot in all history to claim that level of benefit to humanity and turn out to be completely right.

(Bacon didn't singlehandedly invent science, of course, but he did contribute, and may have been the first to realize the power.)

That's the problem with deciding that you'll never admire anything that much: Some ideas really *are* that good. Though no one has *fulfilled* claims more audacious than Bacon's; at least, not yet.

But then how can we resist the happy death spiral[3] with respect to Science itself? The happy death spiral[4] starts when you believe something is *so* wonderful that the halo effect[5] leads you to find *more* and *more* nice things to say about it, making you see it as *even more* wonderful, and so on, spiraling up into the abyss. What if Science is *in fact* so beneficial that we cannot acknowledge its true glory and retain our sanity? Sounds like a nice thing to say, doesn't it? *Oh no it's starting ruuunnnnn...*

1. *http://lesswrong.com/lw/ln/resist_the_happy_death_spiral/*
2. Page 225, 'Affective Death Spirals'.
3. Page 225, 'Affective Death Spirals'.
4. Page 225, 'Affective Death Spirals'.
5. Page 212, 'The Halo Effect'.

If you retrieve the standard[6] cached[7] deep wisdom[8] for *don't go overboard on admiring science,* you will find thoughts like "Science gave us air conditioning, but it also made the hydrogen bomb" or "Science can tell us about stars and biology, but it can never prove or disprove[9] the dragon in my garage[10]." But the people who *originated* such thoughts were *not* trying to resist a happy death spiral. They weren't worrying about their own admiration of science spinning out of control. Probably they didn't like something science had to say about their pet beliefs, and sought ways to undermine its authority.

The *standard* negative things to say about science, aren't likely to appeal to someone who genuinely feels the exultation of science—that's not the intended audience. So we'll have to search for other negative things to say instead.

But if you look selectively for something negative to say about science—even in an attempt to resist a happy death spiral—do you not automatically convict yourself of rationalization[11]? Why would you pay attention to your own thoughts, if you knew you were trying to manipulate yourself[12]?

I am generally skeptical of people who claim that one bias can be used to counteract another. It sounds to me like an automobile mechanic who says that the motor is broken on your right windshield wiper, but instead of fixing it, they'll just break your left windshield wiper to balance things out. This is the sort of cleverness that leads to shooting yourself in the foot. Whatever the solution, it ought to involve believing true things, rather than believing you believe things that you believe are false.

---

6. Page 300, 'The "Outside the Box" Box'.
7. Page 297, 'Cached Thoughts'.
8. Page 314, 'How to Seem (and Be) Deep'.
9. *http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/*
10. Page 43, 'Belief in Belief'.
11. Page 351, 'Rationalization'.
12. Page 401, 'Doublethink (Choosing to be Biased)'.

Can you prevent the happy death spiral by restricting your admiration of Science to a narrow domain? Part of the happy death spiral is seeing the Great Idea everywhere—thinking about how Communism could cure cancer if it was only given a chance. Probably the single most reliable sign of a cult guru is that the guru claims expertise, not in one area, not even in a cluster of related areas, but in *everything*. The guru knows what cult members should eat, wear, do for a living; who they should have sex with; which art they should look at; which music they should listen to...

Unfortunately for this plan, most people fail miserably when they try to describe the neat little box that science has to stay inside. The usual trick, "Hey, science won't cure cancer" isn't going to fly. "Science has nothing to say about a parent's love for their child"—sorry, that's simply false[13]. If you try to sever science from e.g. parental love, you aren't just denying cognitive science and evolutionary psychology. You're also denying Martine Rothblatt's founding of United Therapeutics to seek a cure for her daughter's pulmonary hypertension. (Successfully, I might add.) Science is legitimately related, one way or another, to just about every important facet of human existence.

All right, so what's an example of a *false* nice claim you could make about science?

In my humble opinion, one false claim is that science is so wonderful that scientists shouldn't even try to take ethical responsibility for their work[14], it will automatically end well. This claim, to me, seems to misunderstand the nature of the process whereby science benefits humanity. Scientists are human, they have prosocial concerns just like most other other people, and this is at least *part* of why science ends up doing more good than evil.

13. *http://lesswrong.com/lw/l1/evolutionary_psychology/*
14. *http://intelligence.org/blog/2007/10/21/should-ethicists-be-inside-or-outside-a-profession/*

But that point is, evidently, not beyond dispute. So here's a simpler false nice claim: "A cancer patient can be cured just by publishing enough journal papers." Or, "Sociopaths could become fully normal, if they just committed themselves to never believing anything without replicated experimental evidence with p<0.05."

The way to avoid believing such statements isn't an affective cap, deciding that science is only slightly nice. Nor searching for reasons to believe that publishing journal papers *causes* cancer. Nor believing that science has nothing to say about cancer one way or the other.

Rather, if you know with enough specificity[15] how science works, then you know that, while it may be possible for "science to cure cancer", a cancer patient writing journal papers isn't going to experience a miraculous remission. That *specific* proposed chain of cause and effect is not going to work out.

The happy death spiral is only an emotional problem because of a perceptual problem, the halo effect[16], which makes us more likely to accept future positive claims once we've accepted an initial positive claim. We can't get rid of this effect just by wishing; it will probably always influence us a little. But we can manage to slow down, stop, consider each additional nice claim as an additional burdensome detail[17], and focus on the specific points of the claim apart from its positiveness.

What if a specific nice claim "can't be disproven" but there are arguments "both for and against" it? Actually these are words to be wary of in general, because often this is what people say when they're rehearsing the evidence[18] or avoiding the real weak points[19]. Given the danger of the happy death spiral, it makes sense to try to avoid being happy about *unsettled*

15. Page 58, 'The Virtue of Narrowness'.
16. Page 212, 'The Halo Effect'.
17. *http://lesswrong.com/lw/jk/burdensome_details/*
18. Page 340, 'One Argument Against An Army'.
19. Page 357, 'Avoiding Your Belief's Real Weak Points'.

claims—to avoid making them into a source of yet more positive affect about something you liked already.

The happy death spiral is only a *big* emotional problem because of the overly positive feedback, the ability for the process to go critical. You may not be able to eliminate the halo effect entirely, but you can apply enough critical reasoning to keep the halos subcritical—make sure that the resonance dies out rather than exploding.

You might even say that the whole problem starts with people not bothering to critically examine every additional burdensome detail[20]—demanding sufficient[21] evidence to compensate for complexity[22], searching[23] for flaws as well as support, invoking curiosity[24]—once they've accepted some core premise. Without the conjunction fallacy[25], there might still be a halo effect[26], but there wouldn't be a happy death spiral[27].

Even on the nicest Nice Thingies in the known universe, a perfect rationalist who demanded exactly the necessary evidence for every additional (positive) claim, would experience no affective resonance. You can't do this, but you can stay close enough to rational to keep your happiness from spiraling out of control.

The really dangerous cases are the ones where *any criticism of any positive claim about the Great Thingy feels bad or is socially unacceptable*. Arguments are soldiers, any positive claim is a soldier on our side, stabbing your soldiers in the back is treason.[28] Then the chain reaction goes *super*critical. More on this tomorrow.

---

20. *http://lesswrong.com/lw/jk/burdensome_details/*
21. Page 22, 'How Much Evidence Does It Take?'.
22. Page 29, 'Occam's Razor'.
23. Page 347, 'What Evidence Filtered Evidence?'.
24. Page 478, 'The Meditation on Curiosity'.
25. *http://lesswrong.com/lw/ji/conjunction_fallacy/*
26. Page 212, 'The Halo Effect'.
27. Page 225, 'Affective Death Spirals'.
28. Page 148, 'Politics is the Mind-Killer'.

Addendum: Stuart Armstrong gives closely related advice:[29]

> Cut up your Great Thingy into smaller independent ideas, *and treat them as independent*.
>
> For instance a marxist would cut up Marx's Great Thingy into a theory of value of labour, a theory of the political relations between classes, a theory of wages, a theory on the ultimate political state of mankind. Then each of them should be assessed independently, and the truth or falsity of one should not halo on the others. If we can do that, we should be safe from the spiral, as each theory is too narrow to start a spiral on its own.

This, metaphorically, is like keeping subcritical masses of plutonium from coming together. Three Great Ideas are far less likely to drive you mad than one Great Idea. Armstrong's advice also helps promote specificity: As soon as someone says, "Publishing enough papers can cure your cancer," you ask, "Is that a benefit of the experimental method, and if so, at which stage of the experimental process is the cancer cured? Or is it a benefit of science as a social process, and if so, does it rely on individual scientists wanting to cure cancer, or can they be self-interested?" Hopefully this leads you away from the good or bad feeling, and toward noticing the confusion and lack of support.

Addendum 2: To summarize, you *do* avoid a Happy Death Spiral by (1) splitting the Great Idea into parts (2) treating every additional detail as burdensome (3) thinking about the specifics of the causal chain instead of the good or bad feelings (4) not rehearsing evidence (5) not adding happiness from claims that "you can't *prove* are wrong"; but *not* by (6) refusing to admire anything too much (7) conducting a biased search for

29. *http://lesswrong.com/lw/lm/affective_death_spirals/gp5*

negative points until you feel unhappy again (8) forcibly shoving an idea into a safe box.

9. Uncritical Supercriticality[1]

Followup to: Resist the Happy Death Spiral[2]

Every now and then, you see people arguing over whether atheism is a "religion". As I touched on in Purpose and Pragmatism[3], arguing over the meaning of a word nearly always means that you've lost track of the original question. How might this argument arise to begin with?

An atheist is holding forth, blaming "religion" for the Inquisition, the Crusades, and various conflicts with or within Islam. The religious one may reply, "But atheism is also a religion, because you also have beliefs about God; you believe God doesn't exist." Then the atheist answers, "If atheism is a religion, then not collecting stamps is a hobby," and the argument begins.

Or the one may reply, "But horrors just as great were inflicted by Stalin, who was an atheist, and who suppressed churches in the name of atheism; therefore you are wrong to blame the violence on religion." Now the atheist may be tempted to reply "No true Scotsman[4]", saying, "Stalin's religion was Communism." The religious one answers "If Communism is a religion, then Star Wars fandom is a government," and the argument begins.

Should a "religious" person be defined as someone who has a definite opinion about the existence of at least one God, e.g., assigning a probability lower than 10% or higher than 90% to the existence of Zeus? Or should a "religious" person be defined as someone who has a positive opinion, say a probability higher than 90%, for the existence of at least one God? In the former case, Stalin was "religious"; in the latter case, Stalin was "not religious".

---

1. *http://lesswrong.com/lw/lo/uncritical_supercriticality/*
2. Page 228, 'Resist the Happy Death Spiral'.
3. *http://lesswrong.com/lw/lf/purpose_and_pragmatism/*
4. *http://en.wikipedia.org/wiki/No_true_Scotsman*

But this is exactly the wrong way to look at the problem. What you really want to know—what the argument was originally about—is why, at certain points in human history, large groups of people were slaughtered and tortured, ostensibly in the name of an idea. Redefining a word won't change the facts of history one way or the other.

Communism was a complex catastrophe, and there may be no single *why,* no single critical link in the chain of causality. But if I had to suggest an ur-mistake, it would be... well, I'll let God say it for me:

> "If your brother, the son of your father or of your mother, or your son or daughter, or the spouse whom you embrace, or your most intimate friend, tries to secretly seduce you, saying, 'Let us go and serve other gods,' unknown to you or your ancestors before you, gods of the peoples surrounding you, whether near you or far away, anywhere throughout the world, you must not consent, **you must not listen to him**; you must show him no pity, you must not spare him or conceal his guilt. No, **you must kill him**, your hand must strike the first blow in putting him to death and the hands of the rest of the people following. You must stone him to death, since he has tried to divert you from Yahweh your God." (Deuteronomy 13:7-11, emphasis added)

This was likewise the rule which Stalin set for Communism, and Hitler for Nazism: if your brother tries to tell you why Marx is wrong, if your son tries to tell you the Jews are not planning world conquest, then do not debate him or set forth your own evidence; do not perform replicable experiments or examine history; but turn him in at once to the secret police.

Yesterday, I suggested that one key to resisting[5] an affective death spiral[6] is the principle of "burdensome details[7]"—just *remembering* to question the specific details of each additional nice claim about the Great Idea. (It's not trivial advice. People often don't remember to do this when they're listening to a futurist sketching amazingly detailed projections about the wonders of tomorrow, let alone when they're thinking about their favorite idea ever.) This wouldn't get rid of the halo effect[8], but it would hopefully reduce the resonance to below criticality, so that one nice-sounding claim triggers less than 1.0 additional nice-sounding claims, on average.

The diametric opposite of this advice, which sends the halo effect *super*critical, is when it feels wrong to argue against *any* positive claim about the Great Idea. Politics is the mind-killer[9]. Arguments are soldiers. Once you know which side you're on, you must support all favorable claims, and argue against all unfavorable claims. Otherwise it's like giving aid and comfort to the enemy, or stabbing your friends in the back.

If...

- ...you feel that contradicting someone else who makes a flawed nice claim in favor of evolution[10], would be giving aid and comfort to the creationists;
- ...you feel like you get spiritual credit for each nice thing you say about God, and arguing about it would interfere with your relationship with God;
- ...you have the distinct sense that the other people in the room will dislike you for "not supporting our troops" if you argue against the latest war;
- ...saying anything against Communism gets you ~~stoned to death~~ shot;

5. Page 228, 'Resist the Happy Death Spiral'.
6. Page 225, 'Affective Death Spirals'.
7. *http://lesswrong.com/lw/jk/burdensome_details/*
8. Page 212, 'The Halo Effect'.
9. Page 148, 'Politics is the Mind-Killer'.
10. *http://lesswrong.com/lw/ks/the_wonder_of_evolution/*

...then the affective death spiral has gone supercritical. It is now a Super Happy Death Spiral.

It's not religion, as such, that is the key categorization, relative to our original question: "What makes the slaughter?" The best distinction I've heard[11] between "supernatural" and "naturalistic" worldviews is that a supernatural worldview asserts the existence of ontologically basic mental substances, like spirits, while a naturalistic worldview reduces mental phenomena to nonmental parts. (~~Can't find original source~~ thanks, g![12]) Focusing on this as the source of the problem buys into religious exceptionalism. Supernaturalist claims are worth distinguishing, because they always turn out to be wrong for fairly fundamental[13] reasons. But it's still just one kind of mistake.

An affective death spiral can nucleate around supernatural beliefs; especially monotheisms whose pinnacle is a Super Happy Agent, defined primarily by agreeing with any nice statement about it; especially meme complexes grown sophisticated enough to assert supernatural punishments for disbelief. But the death spiral can also start around a political innovation, a charismatic leader, belief in racial destiny, or an economic hypothesis. The lesson of history is that affective death spirals are dangerous whether or not they happen to involve supernaturalism. Religion isn't special enough, as a class of mistake, to be the key problem.

Sam Harris came closer when he put the accusing finger on *faith*. If you don't place an appropriate burden of proof on each and every additional nice claim, the affective resonance gets started *very* easily. Look at the poor New Agers. Christianity developed defenses against criticism, arguing for the wonders of faith; New Agers culturally inherit the cached thought[14] that faith is positive, but lack Christianity's exclusionary scripture to

11. *http://richardcarrier.blogspot.com/2007/01/defining-supernatural.html*
12. *http://lesswrong.com/lw/lo/uncritical_supercriticality/gqb*
13. Page 96, 'Mysterious Answers to Mysterious Questions'.
14. Page 297, 'Cached Thoughts'.

keep out competing memes. New Agers end up in happy death spirals around stars, trees, magnets, diets, spells, unicorns...

But the affective death spiral turns much deadlier after criticism becomes a sin, or a gaffe, or a crime. There are things in this world that are worth praising greatly, and you can't *flatly* say that praise beyond a certain point is forbidden. But there is *never* an Idea so true that it's wrong to criticize any argument that supports it. Never. Never ever never for ever. *That* is flat. The vast majority[15] of possible beliefs in a nontrivial answer space are false, and likewise, the vast majority of possible *supporting arguments* for a true belief are also false, and not even the happiest idea can change that.

And it is triple ultra forbidden to respond to criticism with violence. There are a very few injunctions in the human art of rationality that have no ifs, ands, buts, or escape clauses. This is one of them. Bad argument gets counterargument. Does not get bullet. Never. Never ever never for ever.

15. Page 22, 'How Much Evidence Does It Take?'.

10. Evaporative Cooling of Group Beliefs[1]

Followup to: Uncritical Supercriticality[2]

Early studiers of cults were surprised to discover than when cults receive a major shock—a prophecy fails to come true, a moral flaw of the founder is revealed—they often come back stronger than before, with increased belief and fanaticism. The Jehovah's Witnesses placed Armageddon in 1975, based on Biblical calculations; 1975 has come and passed. The Unarian cult, still going strong today, survived the nonappearance of an intergalactic spacefleet[3] on September 27, 1975. (The Wikipedia article[4] on Unarianism mentions a failed prophecy in 2001, but makes no mention of the earlier failure in 1975, interestingly enough.)

Why would a group belief become *stronger* after encountering crushing counterevidence?

The conventional interpretation of this phenomenon is based on cognitive dissonance. When people have taken "irrevocable" actions in the service of a belief—given away all their property in anticipation of the saucers landing—they cannot possibly admit they were mistaken. The challenge to their belief presents an immense cognitive dissonance; they must find reinforcing thoughts to counter the shock, and so become more fanatical. In this interpretation, the increased group fanaticism is the result of increased individual fanaticism.

I was looking at a Java applet which demonstrates the use of evaporative cooling to form a Bose-Einstein condensate[5], when it occurred to me that another force entirely might operate to increase fanaticism. Evaporative cooling sets up a potential energy barrier around a collection of hot atoms. Thermal energy

1. *http://lesswrong.com/lw/lr/evaporative_cooling_of_group_beliefs/*
2. Page 235, 'Uncritical Supercriticality'.
3. *http://findarticles.com/p/articles/mi_moSOR/is_n2_v59/ai_20913876/pg_3*
4. *http://en.wikipedia.org/wiki/Unarius_Academy_of_Science*
5. *http://www.colorado.edu/physics/2000/bec/evap_cool.html*

is essentially statistical in nature—not all atoms are moving at the exact same speed. The kinetic energy of any given atom varies as the atoms collide with each other. If you set up a potential energy barrier that's just a little higher than the average thermal energy, the workings of chance will give an occasional atom a kinetic energy high enough to escape the trap. When an unusually fast atom escapes, it takes with an unusually large amount of kinetic energy, and the average energy decreases. The group becomes substantially cooler than the potential energy barrier around it. Playing with the Java applet[6] may make this clearer.

In Festinger's classic "When Prophecy Fails", one of the cult members walked out the door immediately after the flying saucer failed to land. Who gets fed up and leaves *first?* An *average* cult member? Or a relatively more skeptical member, who previously might have been acting as a voice of moderation, a brake on the more fanatic members?

After the members with the highest kinetic energy escape, the remaining discussions will be between the extreme fanatics on one end and the slightly less extreme fanatics on the other end, with the group consensus somewhere in the "middle".

And what would be the analogy to collapsing to form a Bose-Einstein condensate? Well, there's no real need to stretch the analogy that far. But you may recall that I used a fission chain reaction analogy for the affective death spiral; when a group ejects all its voices of moderation, then all the people encouraging each other, and suppressing dissents, may internally increase in average fanaticism. (No thermodynamic analogy here, unless someone develops a nuclear weapon that explodes when it gets cold.)

When Ayn Rand's long-running affair with Nathaniel Branden was revealed to the Objectivist membership, a substantial fraction of the Objectivist membership broke off and followed Branden into espousing an "open system" of Objectivism not

6. *http://www.colorado.edu/physics/2000/bec/evap_cool.html*

bound so tightly to Ayn Rand. Who stayed with Ayn Rand even after the scandal broke? The ones who *really, really* believed in her—and perhaps some of the undecideds, who, after the voices of moderation left, heard arguments from only one side. This may account for how the Ayn Rand Institute is (reportedly) more fanatic after the breakup, than the original core group of Objectivists under Branden and Rand.

A few years back, I was on a transhumanist mailing list where a small group espousing "social democratic transhumanism" vitriolically insulted every libertarian on the list. Most libertarians left the mailing list, most of the others gave up on posting. As a result, the remaining group shifted substantially to the left. Was this deliberate? Probably not, because I don't think the perpetrators knew that much psychology. (For that matter, I can't recall seeing the evaporative cooling analogy elsewhere, though that doesn't mean it hasn't been noted before.) At most, they might have thought to make themselves "bigger fish in a smaller pond".

This is one reason why it's important to be prejudiced in favor of tolerating dissent. Wait until substantially *after* it seems to you justified in ejecting a member from the group, before actually ejecting. If you get rid of the old outliers, the group position will shift, and someone else will become the oddball. If you eject them too, you're well on the way to becoming a Bose-Einstein condensate and, er, exploding.

The flip side: Thomas Kuhn believed that a science has to become a "paradigm", with a shared technical language that excludes outsiders, before it can get any real work done. In the formative stages of a science, according to Kuhn, the adherents go to great pains to make their work comprehensible to outside academics. But (according to Kuhn) a science can only make real progress as a technical discipline once it abandons the requirement of outside accessibility, and scientists working in the paradigm assume familiarity with large cores of technical material in their communications. This sounds cynical, relative to

what is usually said[7] about public understanding of science, but I can definitely see a core of truth here.

My own theory of Internet moderation is that you have to be willing to exclude trolls and spam to get a conversation going. You must even be willing to exclude kindly but technically uninformed folks from technical mailing lists if you want to get any work done. A genuinely open conversation on the Internet degenerates fast. It's the *articulate* trolls that you should be wary of ejecting, on this theory—they serve the hidden function of legitimizing less extreme disagreements. But you should not have so many articulate trolls that they begin arguing with each other, or begin to dominate conversations. If you have one person around who is the famous Guy Who Disagrees With Everything, anyone with a more reasonable, more moderate disagreement won't look like the sole nail sticking out. This theory of Internet moderation may not have served me too well in practice, so take it with a grain of salt.

7. Page 128, 'Applause Lights'.

11. When None Dare Urge Restraint[1]

Followup to: Uncritical Supercriticality[2]

One morning, I got out of bed, turned on my computer, and my Netscape email client automatically downloaded that day's news pane. On that particular day, the news was that two hijacked planes had been flown into the World Trade Center.

These were my first three thoughts, in order:

I guess I really am living in the Future.
Thank goodness it wasn't nuclear.
and then
The overreaction to this will be ten times worse than the original event.

A mere factor of "ten times worse" turned out to be a vast understatement. Even I didn't guess how badly things would go. That's the challenge of pessimism; it's *really hard* to aim low enough that you're pleasantly surprised around as often and as much as you're unpleasantly surprised.

Nonetheless, I did realize immediately that everyone everywhere would be saying how awful, how terrible this event was; and that no one would dare to be the voice of restraint, of proportionate response. Initially, on 9/11, it was thought that six thousand people had died. Any politician who'd said "6000 deaths is 1/8 the annual US casualties from automobile accidents," would have been asked to resign the same hour.

No, 9/11 wasn't a good day. But if *everyone* gets brownie points for emphasizing how much it hurts, and *no one* dares urge restraint in how hard to hit back, then the reaction will be greater than the appropriate level, whatever the appropriate level may be.

1. *http://lesswrong.com/lw/ls/when_none_dare_urge_restraint/*
2. Page 235, 'Uncritical Supercriticality'.

This is the even darker mirror of the happy death spiral[3]—the spiral of hate. Anyone who attacks the Enemy is a patriot; and whoever tries to dissect even a single negative claim about the Enemy is a traitor. But just as the vast majority of all complex statements are untrue, the vast majority of negative things you can say about anyone, even the worst person in the world, are untrue.

I think the best illustration was "the suicide hijackers were cowards[4]". Some common sense, please? It takes a little courage to voluntarily fly your plane into a building. Of all their sins, cowardice was not on the list. But I guess anything bad you say about a terrorist, no matter how silly, must be true. Would I get even more brownie points if I accused al Qaeda of having assassinated John F. Kennedy? Maybe if I accused them of being Stalinists? Really, *cowardice?*

Yes, it matters that the 9/11 hijackers weren't cowards. Not just for understanding the enemy's realistic psychology. There is simply too much damage done by spirals of hate. It is just too dangerous for there to be any target in the world, whether it be the Jews or Adolf Hitler, about whom *saying negative things* trumps *saying accurate things*.

When the defense force contains thousands of aircraft and hundreds of thousands of heavily armed soldiers, one ought to consider that the immune system itself is capable of wreaking more damage than 19 guys and four nonmilitary airplanes. The US spent billions of dollars and thousands of soldiers' lives shooting off its own foot more effectively than any terrorist group could dream.

If the USA had completely ignored the 9/11 attack—just shrugged and rebuilt the building—it would have been better than the real course of history. But that wasn't a political option. Even if anyone privately guessed that the immune response would be more damaging than the disease, American

3. Page 225, 'Affective Death Spirals'.
4. Page 160, 'Are Your Enemies Innately Evil?'.

politicians had no career-preserving choice but to walk straight into al Qaeda's trap. Whoever argues for a greater response is a patriot. Whoever dissects a patriotic claim is a traitor.

Initially, there were smarter responses to 9/11 than I had guessed. I saw a Congressperson—I forget who—say in front of the cameras, "We have forgotten that the first purpose of government is not the economy, it is not health care, it is defending the country from attack." That widened my eyes, that a politician could say something that wasn't an applause light[5]. The emotional shock must have been very great for a Congressperson to say something that... real.

But within two days, the genuine shock faded, and concern-for-image regained total control of the political discourse. Then the spiral of escalation took over completely. Once restraint becomes unspeakable, no matter where the discourse starts out, the level of fury and folly can only rise with time.

Addendum: Welcome[6] redditors! You may also enjoy A Fable of Science and Politics[7] and Policy Debates Should Not Appear One-Sided[8].

5. Page 128, 'Applause Lights'.
6. *http://lesswrong.com/lw/1/about_less_wrong/*
7. Page 143, 'A Fable of Science and Politics'.
8. Page 150, 'Policy Debates Should Not Appear One-Sided'.

12. Every Cause Wants To Be A Cult[1]

Followup to: Correspondence Bias[2], Affective Death Spirals[3], The Robbers Cave Experiment[4]

Cade Metz at *The Register* recently[5] alleged[6] that a secret mailing list of Wikipedia's top administrators has become obsessed with banning all critics and possible critics of Wikipedia. Including banning a productive user when one administrator—solely *because* of the productivity—became convinced that the user was a spy sent by *Wikipedia Review*. And that the top people at Wikipedia closed ranks to defend their own. (I have not investigated these allegations myself, as yet. Hat tip to Eugen Leitl[7].)

Is there some deep moral flaw in seeking to systematize the world's knowledge, which would lead pursuers of that Cause into madness? Perhaps only people with innately totalitarian tendencies would try to become the world's authority on everything—

Correspondence bias[8] alert! (Correspondence bias: making inferences about someone's unique disposition from behavior that can be entirely explained by the situation in which it occurs. When we see someone else kick a vending machine, we think they are "an angry person", but when we kick the vending machine, it's because the bus was late, the train was early and the machine ate our money.) If the allegations about Wikipedia are true, they're explained by *ordinary* human nature, not by *extraordinary* human nature.

---

1. *http://lesswrong.com/lw/lv/every_cause_wants_to_be_a_cult/*
2. Page 157, 'Correspondence Bias'.
3. Page 225, 'Affective Death Spirals'.
4. Page 164, 'The Robbers Cave Experiment'.
5. *http://www.theregister.co.uk/2007/12/04/wikipedia_secret_mailing/*
6. *http://www.theregister.co.uk/2007/12/06/wikipedia_and_overstock/*
7. *http://postbiota.org/pipermail/tt/2007-December/001947.html*
8. Page 157, 'Correspondence Bias'.

The ingroup-outgroup dichotomy[9] is part of ordinary human nature. So are happy death spirals[10] and spirals of hate[11]. A Noble Cause doesn't need a deep hidden flaw for its adherents to form a cultish in-group. It is sufficient that the adherents be human. Everything else follows naturally, decay by default, like food spoiling in a refrigerator after the electricity goes off.

In the same sense that every thermal differential wants to equalize itself, and every computer program wants to become a collection of ad-hoc patches, every Cause *wants* to be a cult. It's a high-entropy state into which the system trends, an attractor in human psychology. It may have nothing to do with whether the Cause is truly Noble. You might think that a Good Cause would rub off its goodness on every aspect of the people associated with it—that the Cause's followers would also be less susceptible to status games, ingroup-outgroup bias, affective spirals, leader-gods. But believing one true idea won't switch off the halo effect[12]. A noble cause won't make its adherents something other than human. There are plenty of bad ideas that can do plenty of damage—but that's not necessarily what's going on.

Every group of people with an unusual goal—good, bad, or silly—will trend toward the cult attractor unless they make a constant effort to resist it. You can keep your house cooler than the outdoors, but you have to run the air conditioner constantly, and as soon as you turn off the electricity—give up the fight against entropy—things will go back to "normal".

On one notable occasion there was a group that went semicultish whose rallying cry was "Rationality! Reason! Objective reality!" (More on this in future posts.) Labeling the Great Idea "rationality" won't protect you any more than putting up a sign over your house that says "Cold!" You still

9. Page 164, 'The Robbers Cave Experiment'.
10. Page 225, 'Affective Death Spirals'.
11. Page 244, 'When None Dare Urge Restraint'.
12. Page 212, 'The Halo Effect'.

have to run the air conditioner—expend the required energy per unit time to reverse the natural slide into cultishness. Worshipping rationality won't make you sane any more than worshipping gravity enables you to fly. You can't talk to thermodynamics and you can't pray to probability theory. You can *use* it, but not join it as an in-group.

Cultishness is quantitative, not qualitative. The question is not "Cultish, yes or no?" but "How much cultishness and where?" Even in Science, which is the archetypal Genuinely Truly Noble Cause, we can readily point to the current frontiers of the war against cult-entropy, where the current battle line creeps forward and back. Are journals more likely to accept articles with a well-known authorial byline, or from an unknown author from a well-known institution, compared to an unknown author from an unknown institution? How much belief is due to authority and how much is from the experiment? Which journals are using blinded reviewers, and how effective is blinded reviewing?

I cite this example, rather than the standard[13] vague accusations of "Scientists aren't open to new ideas", because it shows a *battle line*—a place where human psychology is being actively driven back, where accumulated cult-entropy is being pumped out. (Of course this requires emitting some waste heat.)

This post is not a catalog of techniques for actively pumping against cultishness. Some[14] such[15] techniques I have said before, and some I will say later. *Today* I just want to point out that the worthiness of the Cause does not mean you can spend any *less* effort in resisting the cult attractor. And that if you can point to current battle lines, it does not mean you confess your Noble Cause unworthy. You might think that if the question were "Cultish, yes or no?" that you were obliged to answer "No", or else betray your beloved Cause. But that is like thinking that

13. Page 314, 'How to Seem (and Be) Deep'.
14. Page 228, 'Resist the Happy Death Spiral'.
15. Page 478, 'The Meditation on Curiosity'.

you should divide engines into "perfectly efficient" and "inefficient", instead of measuring waste.

Contrariwise, if you believe that it was the Inherent Impurity of those Foolish Other Causes that made them go wrong, if you laugh at the folly of "cult victims", if you think that cults are led and populated by mutants, then you will not expend the necessary effort to pump against entropy—to resist being human.

13. Guardians of the Truth[1]

Followup to: Tsuyoku Naritai[2], Reversed Stupidity is not Intelligence[3]

The criticism is sometimes leveled against rationalists: "The Inquisition thought *they* had the truth! Clearly this 'truth' business is dangerous."

There are many obvious responses, such as "If you think that possessing the truth *would* license you to torture and kill, you're making a mistake that has nothing to do with epistemology." Or, "So that historical statement you just made about the Inquisition—is it true[4]?"

Reversed stupidity is not intelligence[5]: "If your current computer stops working, you can't conclude that everything about the current system is wrong and that you need a new system without an AMD processor, an ATI video card... even though your current system has all these things and it doesn't work. Maybe you just need a new power cord." To arrive at a poor conclusion requires only one wrong step, not every step wrong. The Inquisitors believed that 2 + 2 = 4, but that wasn't the source of their madness. Maybe epistemological realism wasn't the problem either?

It does seem plausible that if the Inquisition had been made up of relativists, professing that nothing was true and nothing mattered, they would have mustered less enthusiasm for their torture. They would also have had been less enthusiastic if lobotomized. I think that's a fair analogy.

And yet... I think the Inquisition's attitude toward truth played a role. The Inquisition believed that there was such a thing as truth, and that it was important; well, likewise Richard

1. *http://lesswrong.com/lw/lz/guardians_of_the_truth/*
2. *http://lesswrong.com/lw/h8/tsuyoku_naritai_i_want_to_become_stronger/*
3. Page 168, 'Reversed Stupidity Is Not Intelligence'.
4. *http://yudkowsky.net/bayes/truth.html*
5. Page 168, 'Reversed Stupidity Is Not Intelligence'.

Feynman. But the Inquisitors were not Truth-Seekers. They were Truth-*Guardians*.

I once read an argument (can't find source) that a key component of a *zeitgeist* is whether it locates its ideals in its future or its past. Nearly all cultures before the Enlightenment believed in a Fall from Grace—that things had once been perfect in the distant past, but then catastrophe had struck, and everything had slowly run downhill since then:

> "In the age when life on Earth was full... They loved each other and did not know that this was 'love of neighbor'. They deceived no one yet they did not know that they were 'men to be trusted'. They were reliable and did not know that this was 'good faith'. They lived freely together giving and taking, and did not know that they were generous. For this reason their deeds have not been narrated. They made no history."
>
> —*The Way of Chuang Tzu,* trans. Thomas Merton[6]

The perfect age of the past, according to our best anthropological evidence, never existed. But a culture that sees life running inexorably downward is very different from a culture in which you can reach unprecedented heights.

(I say "culture", and not "society", because you can have more than one subculture in a society.)

You could say that the difference between e.g. Richard Feynman and the Inquisition was that the Inquisition believed they *had* truth, while Richard Feynman *sought* truth. This isn't quite defensible, though, because there were undoubtedly some truths that Richard Feynman thought he *had* as well. "The sky is blue," for example, or "2 + 2 = 4".

6. *http://books.google.com/books?id=LDOCZPyg2MQC&pg=PA76&vq=xii&dq=%22chuang+tzu%22+deeds+history&sig=govZRc*

Yes, there are effectively certain truths of science. General Relativity may be overturned by some future physics—albeit not in any way that predicts the Sun will orbit Jupiter; the new theory must steal the successful predictions of the old theory, not contradict them. But evolutionary theory takes place on a higher level of organization than atoms, and nothing we discover about quarks is going to throw out Darwinism, or the cell theory of biology, or the atomic theory of chemistry, or a hundred other brilliant innovations whose truth is now established beyond *reasonable* doubt.

Are these "absolute truths"? Not in the sense of possessing a probability of literally 1.0. But they are cases where science basically thinks it's got the truth.

And yet scientists don't torture people who question the atomic theory of chemistry. Why not? Because they don't believe that certainty licenses torture? Well, yes, that's the *surface* difference; but why *don't* scientists believe this?

Because chemistry asserts no supernatural penalty of eternal torture for disbelieving in the atomic theory of chemistry? But again we recurse and ask the question, "Why?" Why *don't* chemists believe that you go to hell if you disbelieve in the atomic theory?

Because journals won't publish your paper until you get a solid experimental observation of Hell? But all too many scientists can suppress their skeptical reflex at will[7]. Why don't chemists have a private cult which argues that nonchemists go to hell, given that many are Christians anyway?

Questions like that don't have neat single-factor answers. But I would argue that *one* of the factors has to do with assuming a *defensive* posture toward the truth, versus a *productive* posture toward the truth.

When you are the Guardian of the Truth, you've got nothing useful to contribute to the Truth *but* your guardianship of it. When you're trying to win the Nobel Prize in chemistry by

7. *http://lesswrong.com/lw/gv/outside_the_laboratory/*

discovering the next benzene or buckyball, someone who challenges the atomic theory isn't so much a threat to your worldview as a waste of your time.

When you are a Guardian of the Truth, all you can do is try to stave off the inevitable slide into entropy by zapping anything that departs from the Truth. If there's some way to pump against entropy, generate new true beliefs along with a little waste heat, that same pump can keep the truth alive without secret police. In chemistry you can replicate experiments and see for yourself—and that keeps the precious truth alive without need of violence.

And it's not such a terrible threat if we make one mistake somewhere—end up believing a little untruth for a little while—because *tomorrow* we can recover the lost ground.

But this whole trick only works because the experimental method is a "criterion of goodness" which is not a mere "criterion of comparison". Because experiments can recover the truth without need of authority, they can also *override* authority and create new true beliefs where none existed before.

Where there are criteria of goodness that are not criteria of comparison, there can exist *changes* which are *improvements*, rather than *threats*. Where there are *only* criteria of comparison, where there's no way to move past authority, there's also no way to resolve a disagreement between authorities. Except extermination. The bigger guns win.

I don't mean to provide a grand overarching single-factor view of history. I do mean to point out a deep psychological difference between seeing your grand cause in life as *protecting, guarding, preserving*, versus *discovering, creating, improving*. Does the "up" direction of time point to the past or the future? It's a distinction that shades everything, casts tendrils everywhere.

This is why I've always insisted, for example, that if you're going to start talking about "AI ethics", you had better be talking about how you are going to *improve* on the current situation

using AI, rather than just keeping various things from going wrong. Once you adopt criteria of mere comparison, you start losing track of your ideals—lose sight of wrong and right, and start seeing simply "different" and "same".

I would also argue that this basic psychological difference is one of the reasons why an academic field that stops making active progress tends to turn *mean*. (At least by the refined standards of science. *Reputational* assassination is tame by historical standards; most defensive-posture belief systems went for the real thing.) If major shakeups don't arrive often enough to regularly promote young scientists based on merit rather than conformity, the field stops resisting the standard degeneration[8] into authority. When there's not many discoveries being made, there's nothing left to do all day but witch-hunt the heretics.

To get the best mental health benefits of the discover/create/improve posture, you've got to *actually be making progress,* not just hoping for it.

8. Page 247, 'Every Cause Wants To Be A Cult'.

14. Guardians of the Gene Pool[1]

Followup to: Guardians of the Truth[2]

Like any educated denizen of the 21st century, you may have heard of World War II. You may remember that Hitler and the Nazis planned to carry forward a romanticized process of evolution, to breed a new master race, supermen, stronger and smarter than anything that had existed before.

Actually this is a common misconception. Hitler believed that the Aryan superman *had previously existed*—the Nordic stereotype, the blond blue-eyed beast of prey—but had been *polluted* by mingling with impure races. There had been a racial Fall from Grace.

It says something about the degree to which the concept of *progress* permeates Western civilization, that the one is told about Nazi eugenics and hears "They tried to breed a superhuman." *You,* dear reader—if *you* failed hard enough to endorse coercive eugenics, *you* would try to create a superhuman. Because you locate your ideals in your future, not in your past. Because you are *creative.* The thought of breeding back to some Nordic archetype from a thousand years earlier would not even occur to you as a possibility—what, just the *Vikings?* That's *all?* If you failed hard enough to kill, you would damn well try to reach heights never before reached, or what a waste it would all be, eh? Well, that's one reason you're not a Nazi, dear reader.

It says something about how difficult it is for the relatively healthy to envision themselves in the shoes of the relatively sick, that we are told of the Nazis, and distort the tale to make them defective transhumanists.

It's the *Communists* who were the defective transhumanists. "New Soviet Man" and all that. The Nazis were quite definitely the bioconservatives of the tale.

1. *http://lesswrong.com/lw/mo/guardians_of_the_gene_pool/*
2. Page 251, 'Guardians of the Truth'.

15. Guardians of Ayn Rand[1]

Followup to: Every Cause Wants To Be A Cult[2], Guardians of the Truth[3]

> "For skeptics, the idea that reason can lead to a cult is absurd. The characteristics of a cult are 180 degrees out of phase with reason. But as I will demonstrate, not only can it happen, it has happened, and to a group that would have to be considered the unlikeliest cult in history. It is a lesson in what happens when the truth becomes more important than the search for truth..."
> —Michael Shermer, "The Unlikeliest Cult in History[4]"

I think Michael Shermer is over-explaining Objectivism. I'll get around to amplifying on that.

Ayn Rand's novels glorify technology, capitalism, individual defiance of the System, limited government, private property, selfishness[5]. Her ultimate fictional hero, John Galt, was <SPOILER>

a scientist who invented a new form of cheap renewable energy; but then refuses to give it to the world since the profits will only be stolen to prop up corrupt governments.

</SPOILER>

And then—somehow—it all turned into a moral and philosophical "closed system" with Ayn Rand at the center. The term "closed system" is not my own accusation; it's the term the Ayn Rand Institute uses to describe Objectivism. Objectivism

1. *http://lesswrong.com/lw/m1/guardians_of_ayn_rand/*
2. Page 247, 'Every Cause Wants To Be A Cult'.
3. Page 251, 'Guardians of the Truth'.
4. *http://www.2think.org/02_2_she.shtml*
5. *http://lesswrong.com/lw/ky/fake_morality/*

is defined by the works of Ayn Rand. Now that Rand is dead, Objectivism is closed. If you disagree with Rand's works in any respect, you cannot be an Objectivist.

Max Gluckman once said: "A science is any discipline in which the fool of this generation can go beyond the point reached by the genius of the last generation." Science moves forward by slaying its heroes, as Newton fell to Einstein. Every young physicist dreams of being the new champion that future physicists will dream of dethroning.

Ayn Rand's philosophical idol was Aristotle. Now maybe Aristotle was a hot young math talent 2350 years ago, but math has made noticeable progress since his day. Bayesian probability theory is the quantitative logic of which Aristotle's qualitative logic is a special case; but there's no sign that Ayn Rand knew about Bayesian probability theory when she wrote her magnum opus, *Atlas Shrugged*. Rand wrote about "rationality", yet failed to familiarize herself with the modern research in heuristics and biases. How can anyone claim to be a master rationalist, yet know nothing of such elementary subjects?

"Wait a minute," objects the reader, "that's not quite fair! *Atlas Shrugged* was published in 1957! Practically nobody knew about Bayes back then." Bah. Next you'll tell me that Ayn Rand died in 1982, and had no chance to read *Judgment Under Uncertainty: Heuristics and Biases,* which was published that same year.

Science isn't fair. That's sorta the point. An aspiring rationalist in 2007 starts with a huge advantage over an aspiring rationalist in 1957. It's how we know that progress has occurred.

To me the thought of voluntarily embracing a system explicitly tied to the beliefs of one human being, who's *dead*, falls somewhere between the silly and the suicidal. A computer isn't five years old before it's obsolete.

The vibrance that Rand admired in science, in commerce, in every railroad that replaced a horse-and-buggy route, in every

skyscraper built with *new* architecture—it all comes from the principle of *surpassing the ancient masters*. How can there be science, if the most knowledgeable scientist there will ever be, has already lived? Who would raise the New York skyline that Rand admired so, if the tallest building that would ever exist, had already been built?

And yet Ayn Rand acknowledged no superior, in the past, or in the future yet to come. Rand, who began in admiring reason and individuality, ended by ostracizing anyone who dared contradict her. Shermer[6]: "[Barbara] Branden recalled an evening when a friend of Rand's remarked that he enjoyed the music of Richard Strauss. 'When he left at the end of the evening, Ayn said, in a reaction becoming increasingly typical, 'Now I understand why he and I can never be real soulmates. The distance in our sense of life is too great.' Often she did not wait until a friend had left to make such remarks."

Ayn Rand changed over time, one suspects.

Rand grew up in Russia, and witnessed the Bolshevik revolution firsthand. She was granted a visa to visit American relatives at the age of 21, and she never returned. It's easy to hate authoritarianism when you're the victim. It's easy to champion the freedom of the individual, when you are yourself the oppressed.

It takes a much stronger constitution to fear authority when *you* have the power. When people are looking to *you* for answers, it's harder to say "What the hell do I know about music? I'm a writer, not a composer," or "It's hard to see how liking a piece of music can be untrue[7]."

When *you're* the one crushing those who dare offend you, the exercise of power somehow seems much more *justifiable* than when you're the one being crushed. All sorts of excellent justifications[8] somehow leap to mind.

6. *http://www.2think.org/02_2_she.shtml*
7. Page 463, 'Feeling Rational'.
8. Page 371, 'Fake Optimization Criteria'.

Michael Shermer goes into detail on how he thinks that Rand's philosophy ended up descending into cultishness. In particular, Shermer says (it seems) that Objectivism failed because Rand thought that certainty was possible, while science is never certain. I can't back Shermer on that one. The atomic theory of chemistry is pretty damned certain. But chemists haven't become a cult.

Actually, I think Shermer's falling prey to correspondence bias[9] by supposing that there's any particular correlation between Rand's philosophy and the way her followers formed a cult. Every cause wants to be a cult[10].

Ayn Rand fled the Soviet Union, wrote a book about individualism that a lot of people liked, got plenty of compliments, and formed a coterie of admirers. Her admirers found nicer and nicer things to say about her (happy death spiral[11]), and she enjoyed it too much to tell them to shut up. She found herself with the power to crush those of whom she disapproved, and she didn't resist the temptation of power.

Ayn Rand and Nathaniel Branden carried on a secret extramarital affair. (With permission from both their spouses, which counts for a lot in my view. If you want to turn that into a "problem", you have to specify that the spouses were *unhappy*—and then it's still not a matter for outsiders.) When Branden was revealed to have "cheated" on Rand with yet another woman, Rand flew into a fury and excommunicated him. Many Objectivists broke away when news of the affair became public.

Who stayed with Rand, rather than following Branden, or leaving Objectivism altogether? Her *strongest* supporters. Who departed? The previous voices of moderation. (Evaporative cooling of group beliefs.[12]) Ever after, Rand's grip over

9. Page 157, 'Correspondence Bias'.
10. Page 247, 'Every Cause Wants To Be A Cult'.
11. Page 225, 'Affective Death Spirals'.
12. Page 240, 'Evaporative Cooling of Group Beliefs'.

her remaining coterie was absolute, and no questioning was allowed.

The only extraordinary thing about the whole business, is how ordinary it was.

You might think that a belief system which praised "reason" and "rationality" and "individualism" would have gained some kind of special immunity, somehow...?

Well, it didn't.

It worked around as well as putting a sign saying "Cold" on a refrigerator that wasn't plugged in.

The active effort required to resist the slide into entropy wasn't there, and decay inevitably followed.

And if you call that the "unlikeliest cult in history", you're just calling reality nasty names[13].

Let that be a lesson to all of us: Praising[14] "rationality" counts for nothing. Even saying "You must justify your beliefs through Reason, not by agreeing with the Great Leader" just runs a little automatic program that takes whatever the Great Leader says and generates a justification that your fellow followers will view as Reason-able.

So where is the true art of rationality to be found? Studying up on the math of probability theory and decision theory. Absorbing the cognitive sciences like evolutionary psychology, or heuristics and biases. Reading history books...

"Study science, not just me!" is probably the most important piece of advice Ayn Rand should've given her followers and didn't. There's no one human being who ever lived, whose shoulders were broad enough to bear *all* the weight of a true science with many contributors.

It's noteworthy, I think, that Ayn Rand's fictional heroes were architects and engineers; John Galt, her ultimate, was a physicist

13. *http://lesswrong.com/lw/hs/think_like_reality/*
14. Page 84, 'Science as Attire'.

; and yet Ayn Rand herself wasn't a great scientist. As far as I know, she wasn't particularly good at math. She could not aspire to rival her own heroes. Maybe that's why she began to lose track of Tsuyoku Naritai[15].

Now me, y'know, I admire Francis Bacon's audacity[16], but I retain my ability to bashfully confess, "If I could go back in time, and somehow make Francis Bacon understand the problem I'm currently working on[17], his eyeballs would pop out of their sockets like champagne corks and explode."

I admire Newton's accomplishments. But my attitude toward a woman's right to vote, bars me from accepting Newton as a moral paragon. Just as my knowledge of Bayesian probability bars me from viewing Newton as the ultimate unbeatable source of mathematical knowledge. And my knowledge of Special Relativity, paltry and little-used though it may be, bars me from viewing Newton as the ultimate authority on physics.

Newton couldn't realistically have discovered any of the ideas I'm lording over him—*but progress isn't fair! That's the point!*

Science has heroes, but no gods. The great Names are not our superiors, or even our rivals, they are passed milestones on our road; and the most important milestone is the hero yet to come.

To be one more milestone in humanity's road is the best that can be said of anyone; but this seemed too lowly to please Ayn Rand. And that is how she became a mere Ultimate Prophet.

15. *http://lesswrong.com/lw/h8/tsuyoku_naritai_i_want_to_become_stronger/*
16. Page 228, 'Resist the Happy Death Spiral'.
17. *http://intelligence.org/AIRisk.pdf*

16. Two Cult Koans[1]

Followup to: Every Cause Wants To Be A Cult[2]

A novice rationalist studying under the master Ougi was rebuked by a friend who said, "You spend all this time listening to your master, and talking of 'rational' this and 'rational' that—you have fallen into a cult!"

The novice was deeply disturbed; he heard the words, "You have fallen into a cult!" resounding in his ears as he lay in bed that night, and even in his dreams.

The next day, the novice approached Ougi and related the events, and said, "Master, I am constantly consumed by worry that this is all really a cult, and that your teachings are only dogma."

Ougi replied, "If you find a hammer lying in the road and sell it, you may ask a low price or a high one. But if you keep the hammer and use it to drive nails, who can doubt its worth?"

The novice said, "See, now that's just the sort of thing I worry about—your mysterious Zen replies."

Ougi said, "Fine, then, I will speak more plainly, and lay out perfectly reasonable arguments which demonstrate that you have not fallen into a cult. But first you have to wear this silly hat."

Ougi gave the novice a huge brown ten-gallon cowboy hat.

"Er, master..." said the novice.

"When I have explained everything to you," said Ougi, "you will see why this was necessary. Or otherwise, you can continue to lie awake nights, wondering whether this is a cult."

The novice put on the cowboy hat.

Ougi said, "How long will you repeat my words and ignore the meaning? Disordered thoughts begin as feelings of attachment to preferred conclusions. You are too anxious about your self-image as a rationalist. You came to me to seek reassuran-

1. *http://lesswrong.com/lw/m4/two_cult_koans/*
2. Page 247, 'Every Cause Wants To Be A Cult'.

ce. If you had been truly curious[3], not knowing one way or the other, you would have thought of ways to resolve your doubts[4]. Because you needed to resolve your cognitive dissonance, you were willing to put on a silly hat. If I had been an evil man, I could have made you pay a hundred silver coins. When you concentrate on a real-world question, the worth or worthlessness of your understanding will soon become apparent. You are like a swordsman who keeps glancing away to see if anyone might be laughing at him—"

"All *right,*" said the novice.

"You asked for the long version," said Ougi.

This novice later succeeded Ougi and became known as Ni no Tachi. Ever after, he would not allow his students to cite his words in their debates, saying, "Use the techniques and do not mention them."

A novice rationalist approached the master Ougi and said, "Master, I worry that our rationality dojo is... well... a little cultish."

"That is a grave concern," said Ougi.

The novice waited a time, but Ougi said nothing more.

So the novice spoke up again: "I mean, I'm sorry, but having to wear these robes, and the hood—it just seems like we're the bloody Freemasons or something."

"Ah," said Ougi, "the robes and trappings."

"Well, *yes* the robes and trappings," said the novice. "It just seems terribly irrational."

"I will address all your concerns," said the master, "but first you must put on this silly hat." And Ougi drew out a wizard's hat, embroidered with crescents and stars.

The novice took the hat, looked at it, and then burst out in frustration: "*How can this possibly help?*"

"Since you are so concerned about the interactions of clothing

3. Page 478, 'The Meditation on Curiosity'.
4. Page 474, 'The Proper Use of Doubt'.

with probability theory," Ougi said, "it should not surprise you that you must wear a special hat to understand."

When the novice attained the rank of grad student, he took the name Bouzo and would only discuss rationality while wearing a clown suit.

17. Asch's Conformity Experiment[1]

[2] Solomon Asch, with experiments originally carried out in the 1950s and well-replicated since, highlighted a phenomenon now known as "conformity". In the classic experiment, a subject sees a puzzle like the one in the nearby diagram: Which of the lines A, B, and C is the same size as the line X? Take a moment to determine your own answer...

The gotcha is that the subject is seated alongside a number of other people looking at the diagram—seemingly other subjects, actually confederates of the experimenter. The other "subjects" in the experiment, one after the other, say that line C seems to be the same size as X. The real subject is seated next-to-last. How many people, placed in this situation, would say "C"—giving an obviously incorrect answer that agrees with the unanimous answer of the other subjects? What do you think the percentage would be?

Three-quarters of the subjects in Asch's experiment gave a "conforming" answer at least once. A third of the subjects conformed more than half the time.

Interviews after the experiment showed that while most subjects claimed to have not really believed their conforming answers, some said they'd really thought that the conforming option was the correct one.

Asch was disturbed by these results:

> "That we have found the tendency to conformity in our society so strong... is a matter of concern. It

1. *http://lesswrong.com/lw/m9/aschs_conformity_experiment/*
2. *http://scienceaid.co.uk/psychology/social/images/asch.png*

> raises questions about our ways of education and about the values that guide our conduct."

It is not a trivial question whether the subjects of Asch's experiments behaved *irrationally*. Robert Aumann's Agreement Theorem shows that honest Bayesians cannot agree to disagree—if they have common knowledge of their probability estimates, they have the same probability estimate. Aumann's Agreement Theorem was proved more than twenty years after Asch's experiments, but it only formalizes and strengthens an intuitively obvious point—other people's beliefs are often legitimate evidence.

If you were looking at a diagram like the one above, but you knew *for a fact* that the other people in the experiment were honest and seeing the same diagram as you, and three other people said that C was the same size as X, then what are the odds that *only you* are the one who's right? I lay claim to no advantage of *visual* reasoning—I don't think I'm better than an average human at judging whether two lines are the same size. In terms of individual rationality, I hope I would notice my own severe confusion[3] and then assign >50% probability to the majority vote.

In terms of group rationality, seems to me that the proper thing for an honest rationalist to say is, "How surprising, it *looks* to me like B is the same size as X. But if we're all looking at the same diagram and reporting honestly, I have no reason to believe that my assessment is better than yours." The last sentence is important—it's a much weaker claim of disagreement than, "Oh, *I* see the optical illusion—I understand why you think it's C, of course, but the real answer is B."

So the conforming subjects in these experiments are not *automatically* convicted of irrationality, based on what I've described so far. But as you might expect, the devil is in the

3. Page 62, 'Your Strength as a Rationalist'.

details of the experimental results. According to a meta-analysis of over a hundred replications by Smith and Bond (1996):

Conformity increases strongly up to 3 confederates, but doesn't increase further up to 10-15 confederates. If people are conforming rationally, then the opinion of 15 other subjects should be substantially stronger evidence than the opinion of 3 other subjects.

Adding a single dissenter—just one other person who gives the correct answer, or even an incorrect answer that's different from the group's incorrect answer—reduces conformity *very* sharply, down to 5-10%. If you're applying some intuitive version of Aumann's Agreement to think that when 1 person disagrees with 3 people, the 3 are probably right, then in most cases you should be equally willing to think that 2 people will disagree with 6 people. (Not automatically true, but true ceteris paribus.) On the other hand, if you've got people who are emotionally nervous about being the odd one out, then it's easy to see how a single other person who agrees with you, or even a single other person who disagrees with the group, would make you much less nervous.

Unsurprisingly, subjects in the one-dissenter condition did not think their nonconformity had been influenced or enabled by the dissenter. Like the 90% of drivers who think they're ~~above-average~~ in the top 50%, some of them may be right about this, but not all. People are not self-aware of the causes of their conformity or dissent, which weighs against trying to argue them as manifestations of rationality. For example, in the hypothesis that people are socially-rationally choosing to lie in order to not stick out, it appears that (at least some) subjects in the one-dissenter condition do not consciously anticipate the "conscious strategy" they would employ when faced with unanimous opposition.

When the single dissenter suddenly switched to *conforming to the group,* subjects' conformity rates went back up to just as high as in the no-dissenter condition. Being the first dissenter

is a valuable (and costly!) social service, but you've got to keep it up.

Consistently within and across experiments, all-female groups (a female subject alongside female confederates) conform significantly more often than all-male groups. Around one-half the women conform more than half the time, versus a third of the men. If you argue that the average subject is rational, then apparently women are too agreeable and men are too disagreeable, so neither group is actually *rational*...

Ingroup-outgroup manipulations (e.g., a handicapped subject alongside other handicapped subjects) similarly show that conformity is significantly higher among members of an ingroup.

Conformity is lower in the case of blatant diagrams, like the one at the top of this page, versus diagrams where the errors are more subtle. This is hard to explain if (all) the subjects are making a socially rational decision to avoid sticking out.

Added: Paul Crowley reminds me to note that when subjects can respond in a way that will not be seen by the group, conformity also drops, which also argues against an Aumann interpretation.

Asch, S. E. (1956). Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs,* **70.**

Bond, R. and Smith, P. B. (1996.) Culture and Conformity: A Meta-Analysis of Studies Using Asch's (1952b, 1956) Line Judgment Task[4]. *Psychological Bulletin*, **119,** 111-137.

4. *http://www.radford.edu/~jaspelme/_private/gradsoc_articles/individualism_collectivism/conformity%20and%20culture.pdf*

18. Lonely Dissent[1]

Followup to: The Modesty Argument[2], The "Outside the Box" Box[3], Asch's Conformity Experiment[4]

Asch's conformity experiment[5] showed that the presence of a single dissenter tremendously reduced the incidence of "conforming" wrong answers. Individualism is easy, experiment shows, when you have company in your defiance. Every other subject in the room, except one, says that black is white. You become the second person to say that black is black. And it feels glorious: the two of you, lonely and defiant rebels, against the world! (Followup interviews showed that subjects in the one-dissenter condition expressed strong feelings of camaraderie with the dissenter—though, of course, they didn't think the presence of the dissenter had influenced their own nonconformity.)

But you can only *join* the rebellion, after someone, somewhere, becomes the *first* to rebel. Someone has to say that black is black after hearing *everyone* else, one after the other, say that black is white. And that—experiment shows—is a *lot harder*.

Lonely dissent doesn't feel like going to school dressed in black[6]. It feels like going to school wearing a clown suit.

That's the difference between *joining the rebellion* and *leaving the pack*.

If there's one thing I can't stand, it's fakeness—you may have noticed this if you've been reading *Overcoming Bias* for a while. Well, lonely dissent has got to be one of the most

1. *http://lesswrong.com/lw/mb/lonely_dissent/*
2. *http://lesswrong.com/lw/gr/the_modesty_argument/*
3. Page 300, 'The "Outside the Box" Box'.
4. Page 267, 'Asch's Conformity Experiment'.
5. Page 267, 'Asch's Conformity Experiment'.
6. Page 300, 'The "Outside the Box" Box'.

commonly, most ostentatiously faked characteristics around. Everyone wants to be an iconoclast.

I don't mean to degrade the act of joining a rebellion. There are rebellions worth joining. It does take courage to brave the disapproval of your peer group, or perhaps even worse, their shrugs. Needless to say, going to a rock concert is not rebellion. But, for example, vegetarianism is. I'm not a vegetarian myself, but I respect people who are, because I expect it takes a noticeable amount of quiet courage to tell people that hamburgers won't work for dinner. (Albeit that in the Bay Area, people ask as a matter of routine.)

Still, if you tell people that you're a vegetarian, they'll think they understand your motives (even if they don't). They may disagree. They may be offended if you manage to announce it proudly enough, or for that matter, they may be offended just because they're easily offended. But they know how to relate to you.

When someone wears black to school, the teachers and the other children understand the role thereby being assumed in their society. It's Outside the System—in a very standard way that everyone recognizes and understands. Not, y'know, *actually* outside the system. It's a Challenge to Standard Thinking, of a standard sort, so that people indignantly say "I can't understand why you—", but don't have to actually think any thoughts they had not thought before. As the saying goes, "Has any of the 'subversive literature' you've read caused you to modify any of your political views?"

What takes *real* courage is braving the outright *incomprehension* of the people around you, when you do something that *isn't* Standard Rebellion #37, something for which they lack a ready-made script. They don't hate you for a rebel, they just think you're, like, weird, and turn away. This prospect generates a much deeper fear. It's the difference between explaining vegetarianism and explaining cryonics[7]. There are other cry-

7. *http://www.alcor.org/*

onicists in the world, somewhere, but they aren't there next to you. You have to explain it, alone, to people who just think it's *weird*. Not forbidden, but outside bounds that people don't even think about. You're going to get your head frozen? You think that's going to stop you from dying? What do you mean, brain information? Huh? What? Are you *crazy?*

I'm tempted to essay a post facto explanation in evolutionary psychology[8]: You could get together with a small group of friends and walk away from your hunter-gatherer band, but having to go it *alone* in the forests was probably a death sentence—at least reproductively. We don't reason this out explicitly, but that is not the nature of evolutionary psychology. Joining a rebellion that everyone knows about is scary, but nowhere near as scary as doing something really differently. Something that in ancestral times might have ended up, not with the band splitting, but with you being driven out alone.

As the case of cryonics testifies, the fear of thinking *really* different is stronger than the fear of death. Hunter-gatherers had to be ready to face death on a routine basis, hunting large mammals, or just walking around in a world that contained predators. They needed that courage in order to live. Courage to defy the tribe's standard ways of thinking, to entertain thoughts that seem truly weird—well, that probably didn't serve its bearers as well. We don't reason this out explicitly; that's not how evolutionary psychology[9] works. We human beings are just built in such fashion that many more of us go skydiving than sign up for cryonics.

And that's not even the highest courage. There's more than one cryonicist in the world. Only Robert Ettinger had to say it *first*.

To be a *scientific* revolutionary, you've got to be the first person to contradict what everyone else you know is thinking. This is not the only route to scientific greatness; it is rare even

8. *http://lesswrong.com/lw/l1/evolutionary_psychology/*
9. *http://lesswrong.com/lw/l1/evolutionary_psychology/*

among the great. No one can become a scientific revolutionary by trying to imitate revolutionariness. You can only get there by pursuing the correct answer in all things, whether the correct answer is revolutionary or not. But if, in the due course of time—if, having absorbed all the power and wisdom of the knowledge that has already accumulated—if, after all that and a dose of sheer luck, you find your pursuit of mere correctness taking you into new territory... *then* you have an opportunity for your courage to fail.

This is the true courage of lonely dissent, which every damn rock band out there tries to fake.

Of course not everything that takes courage is a good idea. It would take courage to walk off a cliff, but then you would just go splat.

The *fear* of lonely dissent is a hindrance to good ideas, but not every dissenting idea is good. See also Robin Hanson's Against Free Thinkers[10]. Most of the difficulty in having a new true scientific thought is in the "true" part.

It really isn't *necessary* to be different for the sake of being different. If you do things differently only when you see an overwhelmingly good reason, you will have more than enough trouble to last you the rest of your life.

There are a few genuine packs of iconoclasts around. The Church of the SubGenius, for example, seems to genuinely aim at *confusing* the mundanes, not merely offending them. And there are islands of genuine tolerance in the world, such as science fiction conventions. There *are* certain people who have no fear of departing the pack. Many fewer such people really exist, than imagine themselves rebels; but they do exist. And yet scientific revolutionaries are tremendously rarer. Ponder that.

Now *me,* you know, I *really am* an iconoclast. Everyone thinks they are, but with me it's *true,* you see. I would *totally* have worn a clown suit to school. My serious conversations were with books, not with other children.

10. *http://www.overcomingbias.com/2007/06/against_free_th.html*

But if you think you would *totally* wear that clown suit, then don't be too proud of that either! It just means that you need to make an effort in the *opposite direction* to avoid dissenting too easily. That's what I have to do, to correct for my own nature. Other people do have reasons for thinking what they do, and ignoring that completely is as bad as being afraid to contradict them. You wouldn't want to end up as a free thinker[11]. It's not a *virtue,* you see—just a bias either way.

11. *http://www.overcomingbias.com/2007/06/against_free_th.html*

19. Cultish Countercultishness[1]

Followup to: Every Cause Wants To Be A Cult[2], Lonely Dissent[3]

In the modern world, joining a cult is probably one of the worse things that can happen to you. The best-case scenario is that you'll end up in a group of sincere but deluded people, making an honest mistake but otherwise well-behaved, and you'll spend a lot of time and money but end up with nothing to show. Actually, that could describe any failed Silicon Valley startup. Which is supposed to be a hell of a harrowing experience, come to think. So yes, very scary.

Real cults are vastly worse. "Love bombing" as a recruitment technique, targeted at people going through a personal crisis. Sleep deprivation. Induced fatigue from hard labor. Distant communes to isolate the recruit from friends and family. Daily meetings to confess impure thoughts. It's not unusual for cults to take *all* the recruit's money—life savings plus weekly paycheck—forcing them to depend on the cult for food and clothing. Starvation as a punishment for disobedience. Serious brainwashing and serious harm.

With all that taken into account, I should probably sympathize more with people who are terribly nervous, embarking on some odd-seeming endeavor, that *they might be joining a cult*. It should not grate on my nerves. Which it does.

Point one: "Cults" and "non-cults" aren't separated natural kinds like dogs and cats. If you look at any list of cult characteristics[4], you'll see items that could easily describe political parties and corporations—"group members encouraged to distrust outside criticism as having hidden motives", "hierarchical authoritative structure". I've posted on group failure modes

1. *http://lesswrong.com/lw/md/cultish_countercultishness/*
2. Page 247, 'Every Cause Wants To Be A Cult'.
3. Page 271, 'Lonely Dissent'.
4. *http://www.prem-rawat-talk.org/forum/uploads/CultCharacteristics.htm*

like group polarization[5], happy death spirals[6], uncriticality[7], and evaporative cooling[8], all of which seem to feed on each other. When these failures swirl together and meet, they combine to form a Super-Failure stupider than any of the parts, like Voltron[9]. But this is not a cult *essence;* it is a cult *attractor*.

Dogs are born with dog DNA, and cats are born with cat DNA. In the current world, there is no in-between. (Even with genetic manipulation, it wouldn't be as simple as creating an organism with half dog genes and half cat genes.) It's not like there's a mutually reinforcing set of dog-characteristics, which an individual cat can wander halfway into and become a semi-dog.

The human mind, as it thinks about categories, seems to prefer essences to attractors. The one wishes to say "It is a cult" or "It is not a cult", and then the task of classification is over and done. If you observe that Socrates has ten fingers, wears clothes, and speaks fluent Greek, then you can say "Socrates is human" and from there deduce "Socrates is vulnerable to hemlock" without doing specific blood tests to confirm his mortality. You have decided Socrates's humanness once and for all.

But if you observe that a certain group of people seems to exhibit ingroup-outgroup polarization[10] and see a positive halo effect[11] around their Favorite Thing Ever—which could be Objectivism[12], or vegetarianism, or neural networks[13]—you cannot, *from the evidence gathered so far,* deduce whether they have achieved uncriticality[14]. You cannot deduce whether their

5. Page 164, 'The Robbers Cave Experiment'.
6. Page 225, 'Affective Death Spirals'.
7. Page 235, 'Uncritical Supercriticality'.
8. Page 240, 'Evaporative Cooling of Group Beliefs'.
9. *http://www.youtube.com/watch?v=tZZv5Z2Iz_s*
10. Page 164, 'The Robbers Cave Experiment'.
11. Page 212, 'The Halo Effect'.
12. Page 258, 'Guardians of Ayn Rand'.
13. *http://thedailywtf.com/Articles/No,_We_Need_a_Neural_Network.aspx*
14. Page 235, 'Uncritical Supercriticality'.

main idea is true, or false, or genuinely useful but not quite as useful as they think. *From the information gathered so far,* you cannot deduce whether they are otherwise polite, or if they will lure you into isolation and deprive you of sleep and food. The characteristics of cultness are not all present or all absent.

If you look at online arguments over "X is a cult", "X is not a cult", then one side goes through an online list of cult characteristics and finds one that applies and says "Therefore is a cult!" And the defender finds a characteristic that does not apply and says "Therefore it is not a cult!"

You cannot build up an accurate picture of a group's reasoning dynamic using this kind of essentialism. You've got to pay attention to individual characteristics individually.

Furthermore, reversed stupidity is not intelligence[15]. If you're interested in the central *idea,* not just the implementation group, then smart ideas can have stupid followers. Lots of New Agers talk about "quantum physics" but this is no strike against quantum physics. Of course stupid ideas can also have stupid followers. Along with binary essentialism goes the idea that if you infer that a group is a "cult", therefore their beliefs must be false, because false beliefs are characteristic of cults, just like cats have fur. If you're interested in the idea, then look at the idea, not the people[16]. Cultishness is a characteristic of *groups* more than *hypotheses*.

The second error is that when people nervously ask, "This isn't a cult, is it?" it sounds to me like they're seeking *reassurance of rationality*. The notion of a rationalist not getting too attached to their self-image as a rationalist deserves its own post (though see this[17], this[18] and this[19]). But even without going into detail, surely one can see that *nervously seeking reassurance* is not the best frame of mind in which to evaluate

15. Page 168, 'Reversed Stupidity Is Not Intelligence'.
16. Page 178, 'Hug the Query'.
17. *http://yudkowsky.net/virtues/*
18. Page 14, 'Why truth? And...'.
19. Page 264, 'Two Cult Koans'.

questions of rationality. You will not be genuinely curious[20] or think of ways to fulfill your doubts[21]. Instead, you'll find some online source which says that cults use sleep deprivation to control people, you'll notice that Your-Favorite-Group doesn't use sleep deprivation, and you'll conclude "It's not a cult. Whew!" If it doesn't have fur, it must not be a cat. Very reassuring.

But Every Cause Wants To Be A Cult[22], whether the cause itself is wise or foolish. The ingroup-outgroup dichotomy[23] etc. are part of human nature, not a special curse[24] of mutants[25]. Rationality is the exception, not the rule. You have to put forth a constant effort to maintain rationality against the natural slide into entropy. If you decide "It's not a cult!" and sigh with relief, then you will not put forth a continuing effort to push back *ordinary* tendencies toward cultishness. You'll decide the cult-essence is absent, and stop pumping against the entropy of the cult-attractor.

If you are terribly nervous about cultishness, then you will want to deny any hint of any characteristic that resembles a cult. But *any* group with a goal seen in a positive light, is at risk for the halo effect[26], and will have to pump against entropy to avoid an affective death spiral[27]. This is true even for ordinary institutions like political parties—people who think that "liberal values" or "conservative values" can cure cancer, etc. It is true for Silicon Valley startups, both failed and successful. It is true of Mac users and of Linux users. The halo effect[28] doesn't become okay just because everyone does it; if everyone walks off a cliff, you wouldn't too. The error in reasoning is to be fought,

20. Page 478, 'The Meditation on Curiosity'.
21. Page 474, 'The Proper Use of Doubt'.
22. Page 247, 'Every Cause Wants To Be A Cult'.
23. Page 164, 'The Robbers Cave Experiment'.
24. Page 157, 'Correspondence Bias'.
25. Page 160, 'Are Your Enemies Innately Evil?'.
26. Page 212, 'The Halo Effect'.
27. Page 225, 'Affective Death Spirals'.
28. Page 212, 'The Halo Effect'.

not tolerated. But if you're too nervous about "Are you *sure* this isn't a cult?" then you will be reluctant to see *any* sign of cultishness, because that would imply you're in a cult, and *It's not a cult!!* So you won't see the current battlefields where the *ordinary* tendencies toward cultishness are creeping forward, or being pushed back.

The third mistake in nervously asking "This isn't a cult, is it?" is that, I strongly suspect, the *nervousness* is there for entirely the wrong reasons.

Why is it that groups which praise their Happy Thing to the stars, encourage members to donate all their money and work in voluntary servitude, and run private compounds in which members are kept tightly secluded, are called "religions" rather than "cults" once they've been around for a few hundred years?

Why is it that most of the people who nervously ask of cryonics, "This isn't a cult, is it?" would not be equally nervous about attending a Republican or Democrat political rally? Ingroup-outgroup dichotomies[29] and happy death spirals[30] can happen in political discussion, in mainstream religions, in sports fandom. If the *nervousness* came from fear of *rationality errors,* people would ask "This isn't an ingroup-outgroup dichotomy[31], is it?" about Democrat or Republican political rallies, in just the same fearful tones.

There's a legitimate reason to be less fearful of Libertarianism than of a flying-saucer cult, because Libertarians don't have a reputation for employing sleep deprivation to convert people. But cryonicists don't have a reputation for using sleep deprivation, either. So why be any more worried about having your head frozen after you stop breathing[32]?

I suspect that the *nervousness* is not the fear of believing falsely, or the fear of physical harm. It is the fear of lonely

29. Page 164, 'The Robbers Cave Experiment'.
30. Page 225, 'Affective Death Spirals'.
31. Page 164, 'The Robbers Cave Experiment'.
32. *http://www.alcor.org/*

dissent[33]. The nervous feeling that subjects get in Asch's conformity experiment[34], when all the other subjects (actually confederates) say one after another that line C is the same size as line X, and it looks to the subject like line B is the same size as line X. The fear of leaving the pack.

That's why groups whose beliefs have been around long enough to seem "normal" don't inspire the same nervousness as "cults", though some mainstream religions may also take all your money and send you to a monastery. It's why groups like political parties, that are strongly liable for rationality errors, don't inspire the same nervousness as "cults". The word "cult" isn't being used to symbolize rationality errors, it's being used as a label for something that *seems weird.*

Not every change is an improvement, but every improvement is necessarily a change. That which you want to do better, you have no choice but to do differently. Common wisdom does embody a fair amount of, well, actual wisdom; yes, it makes sense to require an extra burden of proof for weirdness. But the *nervousness* isn't that kind of deliberate, rational consideration. It's the fear of believing something that will make your friends look at you really oddly. And so people ask "This isn't a *cult,* is it?" in a tone that they would never use for attending a political rally, or for putting up a gigantic Christmas display.

That's the part that bugs me.

It's as if, as soon as you believe anything that your ancestors did not believe, the Cult Fairy comes down from the sky and infuses you with the Essence of Cultness, and the next thing you know, you're all wearing robes[35] and chanting[36]. As if "weird" beliefs are the *direct cause* of the problems, never mind the sleep deprivation and beatings. The harm done by cults—the Heaven's Gate suicide and so on—just goes to show that every-

33. Page 271, 'Lonely Dissent'.
34. Page 267, 'Asch's Conformity Experiment'.
35. Page 264, 'Two Cult Koans'.
36. Page 184, 'The Litany Against Gurus'.

one with an odd belief is crazy; the first and foremost characteristic of "cult members" is that they are Outsiders with Peculiar Ways.

Yes, socially unusual belief puts a group at risk for ingroup-outgroup thinking[37] and evaporative cooling[38] and other problems. But the unusualness is a risk factor, not a disease in itself. Same thing with having a goal that you think is worth accomplishing. Whether or not the belief is true, having a nice goal always puts you at risk of the happy death spiral[39]. But that makes lofty goals a risk factor, not a disease. Some goals are genuinely worth pursuing[40].

On the other hand, I see no legitimate reason for sleep deprivation or threatening dissenters with beating, full stop[41]. When a group does this, then whether you call it "cult" or "not-cult", you have directly answered[42] the pragmatic question of whether to join.

Problem four: The fear of lonely dissent is something that *cults themselves* exploit. Being afraid of your friends looking at you disapprovingly is *exactly the effect that real cults use to convert and keep members*—surrounding converts with wall-to-wall agreement among cult believers.

The fear of strange ideas, the impulse to conformity[43], has no doubt warned many potential victims away from flying-saucer cults. When you're out, it keeps you out. But when you're *in,* it keeps you *in*. Conformity just glues you to wherever you are, whether that's a good place or a bad place.

The one wishes there was some way they could be *sure* that they weren't in a "cult". Some definite, crushing rejoinder to

37. Page 164, 'The Robbers Cave Experiment'.
38. Page 240, 'Evaporative Cooling of Group Beliefs'.
39. Page 225, 'Affective Death Spirals'.
40. Page 228, 'Resist the Happy Death Spiral'.
41. Page 235, 'Uncritical Supercriticality'.
42. Page 178, 'Hug the Query'.
43. Page 267, 'Asch's Conformity Experiment'.

people who looked at them funny. Some way they could know once and for all that they were doing the right thing, without these constant doubts. I believe that's called "need for closure". And—of course—cults exploit that, too.

Hence the phrase, "Cultish countercultishness."

Living with doubt is not a virtue—the purpose of every doubt is to annihilate itself[44] in success or failure, and a doubt that just hangs around, accomplishes nothing. But sometimes a doubt does take a while to annihilate itself. Living with a stack of currently unresolved doubts is an unavoidable fact of life for rationalists. Doubt shouldn't be scary. Otherwise you're going to have to choose between living one heck of a hunted life, or one heck of a stupid one.

If you really, genuinely can't figure out whether a group is a "cult", then you'll just have to choose under conditions of uncertainty. That's what decision theory is all about.

Problem five: Lack of strategic thinking.

I know people who are cautious around Singularitarianism[45], and they're *also* cautious around political parties and mainstream religions. *Cautious,* not nervous or defensive. These people can see at a glance that Singularitarianism is obviously not a full-blown cult with sleep deprivation etc. But they worry that Singularitarianism will *become* a cult, because of risk factors like turning the concept of a powerful AI into a Super Happy Agent[46] (an agent defined primarily by agreeing with any nice thing said about it). Just because something isn't a cult now, doesn't mean it won't become a cult in the future. Cultishness is an attractor, not an essence.

Does *this* kind of caution annoy me? Hell no. I spend a lot of time worrying about that scenario myself. I try to place my Go stones in advance to block movement in that direction.

44. Page 474, 'The Proper Use of Doubt'.
45. *http://intelligence.org/AIRisk.pdf*
46. Page 235, 'Uncritical Supercriticality'.

Hence, for example, the series of posts on cultish failures of reasoning.

People who talk about "rationality" also have an added risk factor. Giving people advice about how to think is an inherently dangerous business. But it is a *risk factor,* not a *disease.*

Both of my favorite Causes are at-risk for cultishness. Yet somehow, I get asked "Are you sure this isn't a cult?" a lot more often when I talk about powerful AIs, than when I talk about probability theory and cognitive science. I don't know if one risk factor is higher than the other, but I know which one *sounds weirder*...

Problem #6 with asking "This isn't a cult, is it?"...

Just the question itself places me in a very annoying sort of Catch-22. An actual Evil Guru would surely use the one's nervousness against them, and design a plausible elaborate argument explaining Why This Is Not A Cult, and the one would be eager to accept it. Sometimes I get the impression that this is what people *want* me to do! Whenever I try to write about cultishness and how to avoid it, I keep feeling like I'm giving in to that flawed desire—that I am, in the end, providing people with *reassurance.* Even when I tell people that a constant fight against entropy is required.

It feels like I'm making myself a first dissenter in Asch's conformity experiment, telling people, "Yes, line X really is the same as line B, it's okay for you to say so too." They shouldn't need to ask! Or, even worse, it feels like I'm presenting an elaborate argument for Why This Is Not A Cult. It's a *wrong question.*

Just look at the group's reasoning processes for yourself, and decide for yourself whether it's something you want to be part of, once you get rid of the fear of weirdness. It is your own responsibility to stop yourself from thinking cultishly, no matter which group you currently happen to be operating in.

Once someone asks "This isn't a cult, is it?" then no matter how I answer, I always feel like I'm defending something. I do

not like this feeling. It is not the function of a Bayesian Master[47] to give reassurance, nor of rationalists to defend.

Cults feed on groupthink, nervousness, desire for reassurance. You cannot make nervousness go away by wishing, and false self-confidence is even worse. But so long as someone needs reassurance—even reassurance about being a rationalist—that will always be a flaw in their armor. A skillful swordsman focuses on the target[48], rather than glancing away to see if anyone might be laughing. When you know what you're trying to do and why, you'll know whether you're getting it done or not, and whether a group is helping you or hindering you.

(PS: If the one comes to you and says, "Are you *sure* this isn't a cult?", don't try to explain all these concepts in one breath. You're underestimating inferential distances[49]. The one will say, "Aha, so you're *admitting* you're a cult!" or "Wait, you're saying I shouldn't worry about joining cults?" or "So... the fear of cults is cultish? That sounds awfully cultish to me." So the last annoyance factor—#7 if you're keeping count—is that all of this is such a long story to explain.)

47. Page 264, 'Two Cult Koans'.
48. Page 178, 'Hug the Query'.
49. *http://lesswrong.com/lw/kg/expecting_short_inferential_distances/*

Seeing with Fresh Eyes

A sequence on the incredibly difficult feat of getting your brain to actually think about something, instead of instantly stopping on the first thought that comes to mind.

This is sometimes referred to as "thinking outside the box" by people who, for your convenience, will go on to helpfully point out exactly where "outside the box" is located. The Less Wrong version is called "thinking outside the 'Outside the Box' box". Isn't it funny how nonconformists all dress the same...

1. Anchoring and Adjustment[1]

Suppose I spin a Wheel of Fortune device as you watch, and it comes up pointing to 65. Then I ask: Do you think the percentage of African countries in the UN is above or below this number? What do you think is the percentage of African countries in the UN? Take a moment to consider these two questions yourself, if you like, and please don't Google.

Also, try to guess, within *5 seconds,* the value of the following arithmetical expression. 5 seconds. Ready? Set... *Go!*

1 x 2 x 3 x 4 x 5 x 6 x 7 x 8

Tversky and Kahneman (1974) recorded the estimates of subjects who saw the Wheel of Fortune showing various numbers. The median estimate of subjects who saw the wheel show 65 was 45%; the median estimate of subjects who saw 10 was 25%.

The current theory for this and similar experiments is that subjects take the initial, uninformative number as their starting point or *anchor;* and then they *adjust* upward or downward from their starting estimate until they reached an answer that "sounded plausible"; and then they stopped adjusting. This typically results in under-adjustment from the anchor—more distant numbers could also be "plausible", but one stops at the first satisfying-sounding answer.

Similarly, students shown "1 x 2 x 3 x 4 x 5 x 6 x 7 x 8" made a median estimate of 512, while students shown "8 x 7 x 6 x 5 x 4 x 3 x 2 x 1" made a median estimate of 2,250. The motivating hypothesis was that students would try to multiply (or guess-combine) the first few factors of the product, then adjust upward. In both cases the adjustments were insufficient, relative to the true value of 40,320; but the first set of guesses were much more insufficient because they started from a lower anchor.

1. *http://lesswrong.com/lw/j7/anchoring_and_adjustment/*

Tversky and Kahneman report that offering payoffs for accuracy did not reduce the anchoring effect.

Strack and Mussweiler (1997) asked for the year Einstein first visited the United States. Completely implausible anchors, such as 1215 or 1992, produced anchoring effects just as large as more plausible anchors such as 1905 or 1939.

There are obvious applications in, say, salary negotiations, or buying a car. I won't suggest that you exploit it, but watch out for exploiters.

And: Watch yourself thinking, and try to notice when you are *adjusting* a figure in search of an estimate.

Debiasing manipulations for anchoring have generally proved not very effective. I would suggest these two: First, if the initial guess sounds implausible, try to throw it away entirely and come up with a new estimate, rather than sliding from the anchor. But this in itself may not be sufficient—subjects instructed to avoid anchoring still seem to do so (Quattrone et. al. 1981). So second, even if you are trying the first method, try also to think of an anchor in the opposite direction—an anchor that is clearly too small or too large, instead of too large or too small—and dwell on it briefly.

Quattrone, G.A., Lawrence, C.P., Finkel, S.E., & Andrus, D.C. (1981). Explorations in anchoring: The effects of prior range, anchor extremity, and suggestive hints. Manuscript, Stanford University.

Strack, F. & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. Journal of Personality and Social Psychology, 73, 437-446.

Tversky, A. and Kahneman, D. 1974. Judgment under uncertainty: Heuristics and biases. Science, 185:1124--1131.

2. Priming and Contamination[1]

Suppose you ask subjects to press one button if a string of letters forms a word, and another button if the string does not form a word. (E.g., "banack" vs. "banner".) Then you show them the string "water". Later, they will more quickly identify the string "drink" as a word. This is known as "cognitive priming"; this particular form would be "semantic priming" or "conceptual priming".

The fascinating thing about priming is that it occurs at such a low level—priming speeds up *identifying letters as forming a word,* which one would expect to take place *before* you deliberate on the word's meaning.

Priming also reveals the massive parallelism of spreading activation: if seeing "water" activates the word "drink", it probably also activates "river", or "cup", or "splash"... and this activation spreads, from the semantic linkage of concepts, all the way back to recognizing strings of letters.

Priming is subconscious and unstoppable, an artifact of the human neural architecture. Trying to stop yourself from priming is like trying to stop the spreading activation of your own neural circuits. Try to say aloud the color—not the meaning, but the color—of the following letter-string: "GREEN"

In Mussweiler and Strack (2000), subjects were asked the anchoring question[2]: "Is the annual mean temperature in Germany higher or lower than 5 Celsius / 20 Celsius?" Afterward, on a word-identification task, subjects presented with the 5 Celsius anchor were faster on identifying words like "cold" and "snow", while subjects with the high anchor were faster to identify "hot" and "sun". This shows a non-adjustment mechanism for anchoring: priming compatible thoughts and memories.

The more general result is that *completely uninformative, known false,* or *totally irrelevant* "information" can influence

1. *http://lesswrong.com/lw/k3/priming_and_contamination/*
2. Page 289, 'Anchoring and Adjustment'.

estimates and decisions. In the field of heuristics and biases, this more general phenomenon is known as *contamination*. (Chapman and Johnson 2002.)

Early research in heuristics and biases discovered anchoring effects[3], such as subjects giving lower (higher) estimates of the percentage of UN countries found within Africa, depending on whether they were first asked if the percentage was more or less than 10 (65). This effect was originally attributed to subjects adjusting from the anchor as a starting point, stopping as soon as they reached a plausible value, and under-adjusting because they were stopping at one end of a confidence interval. (Tversky and Kahneman 1974.)

Tversky and Kahneman's early hypothesis still appears to be the correct explanation in some circumstances, notably when subjects generate the initial estimate themselves (Epley and Gilovich 2001). But modern research seems to show that most anchoring is actually due to contamination, not sliding adjustment. (Hat tip for Unnamed[4] for reminding me of this—I'd read the Epley/Gilovich paper years ago, as a chapter in *Heuristics and Biases,* but forgotten it.)

Your grocery store probably has annoying signs saying "Limit 12 per customer" or "5 for $10". Are these signs effective at getting customers to buy in larger quantities? You probably think you're not influenced[5]. But *someone* must be, because these signs have been shown to work, which is why stores keep putting them up. (Wansink et. al. 1998.)

Yet the most fearsome aspect of contamination is that it serves as yet[6] another[7] of[8] the[9] thousand[10] faces[11] of[12] confir-

3. Page 289, 'Anchoring and Adjustment'.
4. *http://lesswrong.com/lw/j7/anchoring_and_adjustment/eza*
5. *http://en.wikipedia.org/wiki/Bias_blind_spot*
6. Page 108, 'Positive Bias: Look Into the Dark'.
7. Page 333, 'Knowing About Biases Can Hurt People'.
8. Page 426, 'The Third Alternative'.
9. Page 71, 'Hindsight bias'.
10. Page 318, 'We Change Our Minds Less Often Than We Think'.

mation[13] bias[14]. Once an idea gets into your head, it primes information compatible with it—and thereby ensures its continued existence. Never mind the selection pressures for winning political arguments; confirmation bias is built directly into our hardware, associational networks priming compatible thoughts and memories. An unfortunate side effect of our existence as neural creatures.

A single fleeting image can be enough to prime associated words for recognition. Don't think it takes anything more to set confirmation bias in motion. All it takes is that one quick flash, and the bottom line is already decided[15], for we change our minds less often than we think[16]...

Chapman, G.B. and Johnson, E.J. 2002. Incorporating the irrelevant: Anchors in judgments of belief and value[17]. In Gilovich et. al. (2003).

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science*, **12,** 391–396.

Mussweiler, T. and Strack, F. Comparing is believing: a selective accessibility model of judgmental anchoring. *European Review of Social Psychology,* **10**, 135-167.

Tversky, A. and Kahneman, D. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, **185**: 251-284.

Wansink, B., Kent, R.J. and Hoch, S.J. 1998. An Anchoring and Adjustment Model of Purchase Quantity Decisions. *Journal of Marketing Research*, **35**(February): 71-81.

11. Page 87, 'Fake Causality'.
12. Page 340, 'One Argument Against An Army'.
13. Page 351, 'Rationalization'.
14. Page 347, 'What Evidence Filtered Evidence?'.
15. Page 343, 'The Bottom Line'.
16. Page 318, 'We Change Our Minds Less Often Than We Think'.
17. *http://web.archive.org/web/20040325202602/http://cebiz.org/ejj/PDF%20Papers/Incorporating%20the%20Irrelevant.PDF*

3. Do We Believe Everything We're Told?[1]

Some early experiments on anchoring[2] and adjustment[3] tested whether *distracting* the subjects—rendering subjects cognitively "busy" by asking them to keep a lookout for "5" in strings of numbers, or some such—would decrease adjustment, and hence increase the influence of anchors. Most of the experiments seemed to bear out the idea that cognitive busyness increased anchoring, and more generally contamination[4].

Looking over the accumulating experimental results—more and more findings of contamination, exacerbated by cognitive busyness—Daniel Gilbert saw a truly crazy pattern emerging: Do we believe *everything* we're told?

One might naturally think that on being told a proposition, we would first *comprehend* what the proposition meant, then *consider* the proposition, and finally *accept* or *reject* it. This obvious-seeming model of cognitive process flow dates back to Descartes. But Descartes's rival, Spinoza, disagreed; Spinoza suggested that we first *passively accept a proposition in the course of comprehending it,* and only afterward *actively disbelieve* propositions which are rejected by consideration.

Over the last few centuries, philosophers pretty much went along with Descartes, since his view seemed more, y'know, logical and intuitive[5]. But Gilbert saw a way of testing Descartes's and Spinoza's hypotheses experimentally.

If Descartes is right, then distracting subjects should interfere with both accepting true statements and rejecting false statements. If Spinoza is right, then distracting subjects should cause them to remember false statements as being true, but

1. *http://lesswrong.com/lw/k4/do_we_believe_everything_were_told/*
2. Page 291, 'Priming and Contamination'.
3. Page 289, 'Anchoring and Adjustment'.
4. Page 291, 'Priming and Contamination'.
5. *http://www.overcomingbias.com/2007/10/what-evidence-i.html*

should not cause them to remember true statements as being false.

Gilbert, Krull, and Malone[6] (1990) bears out this result, showing that, among subjects presented with novel statements labeled TRUE or FALSE, distraction had no effect on identifying true propositions (55% success for uninterrupted presentations, vs. 58% when interrupted); but did affect identifying false propositions (55% success when uninterrupted, vs. 35% when interrupted).

A much more dramatic illustration was produced in followup experiments by Gilbert, Tafarodi and Malone[7] (1993). Subjects read aloud crime reports crawling across a video monitor, in which the color of the text indicated whether a particular statement was true or false. Some reports contained false statements that exacerbated the severity of the crime, other reports contained false statements that extenuated (excused) the crime. Some subjects also had to pay attention to strings of digits, looking for a "5", while reading the crime reports—this being the distraction task to create cognitive busyness. Finally, subjects had to recommend the length of prison terms for each criminal, from 0 to 20 years.

Subjects in the cognitively busy condition recommended an average of 11.15 years in prison for criminals in the "exacerbating" condition, that is, criminals whose reports contained labeled false statements exacerbating the severity of the crime. Busy subjects recommended an average of 5.83 years in prison for criminals whose reports contained labeled false statements excusing the crime. This nearly twofold difference was, as you might suspect, statistically significant.

Non-busy participants read exactly the same reports, with the same labels, and the same strings of numbers occasionally crawling past, except that they did not have to search for the

6. *http://www.wjh.harvard.edu/%7Edtg/Gilbert%20et%20al%20%28UNBELIEVING%29.pdf*
7. *http://www.wjh.harvard.edu/~dtg/Gilbert%20et%20al%20(EVERYTHING%20YOU%20READ).pdf*

number "5". Thus, they could devote more attention to "unbelieving" statements labeled false. These non-busy participants recommended 7.03 years versus 6.03 years for criminals whose reports falsely exacerbated or falsely excused.

Gilbert, Tafarodi and Malone's paper was entitled "You Can't Not Believe Everything You Read".

This suggests —to say the very least—that we should be more careful when we expose ourselves to unreliable information, especially if we're doing something else at the time. Be careful when you glance at that newspaper in the supermarket.

PS: According to an unverified rumor I just made up, people will be less skeptical of this blog post because of the distracting color changes.

Gilbert, D. 2002. Inferential correction. In *Heuristics and biases: The psychology of intuitive judgment*. You recognize this citation by now, right?

Gilbert, D., Krull, D. and Malone, P. 1990. Unbelieving the unbelievable: Some problems in the rejection of false information.[8] *Journal of Personality and Social PSychology,* **59**(4), 601-613.

Gilbert, D., Tafarodi, R. and Malone, P. 1993. You can't not believe everything you read.[9] *Journal of Personality and Social Psychology,* **65**(2), 221-233.

8. *http://www.wjh.harvard.edu/%7Edtg/Gilbert%20et%20al%20%28UNBELIEVING%29.pdf*
9. *http://www.wjh.harvard.edu/~dtg/Gilbert%20et%20al%20(EVERYTHING%20YOU%20READ).pdf*

4. Cached Thoughts[1]

One of the single greatest puzzles about the human brain is how the damn thing works *at all* when most neurons fire 10-20 times per second, or 200Hz tops. In neurology, the "hundred-step rule" is that any postulated operation has to complete in *at most* 100 sequential steps—you can be as parallel as you like, but you can't postulate more than 100 (preferably less) neural spikes one after the other.

Can you imagine having to program using 100Hz CPUs, no matter how many of them you had? You'd also need a hundred billion processors just to get *anything* done in realtime.

If you did need to write realtime programs for a hundred billion 100Hz processors, one trick you'd use as heavily as possible is caching. That's when you store the results of previous operations and look them up next time, instead of recomputing them from scratch. And it's a very *neural* idiom—recognition, association, completing the pattern.

It's a good guess that the actual *majority* of human cognition consists of cache lookups.

This thought does tend to go through my mind at certain times.

There was a wonderfully illustrative story which I thought I had bookmarked, but couldn't re-find: it was the story of a man whose know-it-all neighbor had once claimed in passing that the best way to remove a chimney from your house was to knock out the fireplace, wait for the bricks to drop down one level, knock out those bricks, and repeat until the chimney was gone. Years later, when the man wanted to remove his own chimney, this cached thought was lurking, waiting to pounce...

As the man noted afterward—you can guess it didn't go well—his neighbor was not particularly knowledgeable in these matters, not a trusted source. If he'd *questioned* the idea, he

1. *http://lesswrong.com/lw/k5/cached_thoughts/*

probably would have realized it was a poor one. Some cache hits we'd be better off recomputing. But the brain completes the pattern automatically—and if you don't consciously realize the pattern needs correction, you'll be left with a completed pattern.

I suspect that if the thought had occurred to the man himself—if he'd *personally* had this bright idea for how to remove a chimney—he would have examined the idea more critically. But if someone *else* has already thought an idea through, you can save on computing power by caching their *conclusion*—right?

In modern civilization particularly, no one can think fast enough to think their own thoughts. If I'd been abandoned in the woods as an infant, raised by wolves or silent robots, I would scarcely be recognizable as human. No one can think fast enough to recapitulate the wisdom of a hunter-gatherer tribe in one lifetime, starting from scratch. As for the wisdom of a literate civilization, forget it.

But the flip side of this is that I continually see people who aspire to critical thinking, repeating back cached thoughts which were not invented by critical thinkers.

A good example is the skeptic who concedes, "Well, you can't prove or disprove a religion by factual evidence." As I have pointed out elsewhere[2], this is simply false as probability theory. And it is also simply false relative to the real psychology of religion—a few centuries ago, saying this would have gotten you burned at the stake. A mother whose daughter has cancer prays, "God, please heal my daughter", not, "Dear God, I know that religions are not allowed to have any falsifiable consequences, which means that you can't possibly heal my daughter, so... well, basically, I'm praying to make myself feel better, instead of doing something that could actually help my daughter."

But people read "You can't prove or disprove a religion by factual evidence," and then, the next time they see a piece of ev-

2. *http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/*

idence disproving a religion, their brain completes the pattern. Even some atheists repeat this absurdity without hesitation. If they'd thought of the idea themselves, rather than hearing it from someone else, they would have been more skeptical.

Death: complete the pattern: "Death gives meaning to life."

It's frustrating, talking to good and decent folk—people who would never in a thousand years *spontaneously* think of wiping out the human species—raising the topic of existential risk, and hearing them say, "Well, maybe the human species doesn't deserve to survive." They would never in a thousand years shoot their own child, who is a part of the human species, but the brain completes the pattern.

What patterns are being completed, inside your mind, that you never chose to be there?

Rationality: complete the pattern: "Love isn't rational."

If this idea had suddenly occurred to you personally, as an entirely new thought, how would you examine it critically? I know what *I* would say[3], but what would *you?* It can be hard to see with fresh eyes. Try to keep your mind from completing the pattern in the standard, unsurprising, already-known way. It may be that there is no better answer than the standard one, but you can't *think* about the answer until you can stop your brain from filling in the answer automatically.

Now that you've read this blog post, the next time you hear someone unhesitatingly repeating a meme you think is silly or false, you'll think, "Cached thoughts." My belief is now there in your mind, waiting to complete the pattern. But is it true? Don't let your mind complete the pattern! *Think!*

3. Page 463, 'Feeling Rational'.

5. The "Outside the Box" Box[1]

Whenever someone exhorts you to "think outside the box", they usually, *for your convenience,* point out exactly where "outside the box" is located. Isn't it funny how nonconformists all dress the same...

In Artificial Intelligence, everyone outside the field has a cached result[2] for *brilliant new revolutionary AI idea*—neural networks, which work just like the human brain! New AI Idea: complete the pattern: "Logical AIs, despite all the big promises, have failed to provide real intelligence for decades—what we need are neural networks!"

This cached thought has been around for three decades. Still no general intelligence. But, somehow, everyone outside the field knows that neural networks are the Dominant-Paradigm-Overthrowing New Idea, ever since backpropagation was invented in the 1970s. Talk about your aging hippies.

Nonconformist images, by their nature, permit no departure from the norm. If you don't wear black, how will people know you're a tortured artist? How will people recognize uniqueness if you don't fit the standard pattern for what uniqueness is supposed to look like? How will anyone recognize you've got a revolutionary AI concept, if it's not about neural networks?

Another example of the same trope is "subversive" literature, all of which sounds the same, backed up by a tiny defiant league of rebels who control the entire English Department. As Anonymous asks[3] on Scott Aaronson's blog:

> "Has any of the subversive literature you've read caused you to modify any of your political views?"

Or as Lizard observes[4]:

1. *http://lesswrong.com/lw/k6/the_outside_the_box_box/*
2. Page 297, 'Cached Thoughts'.
3. *http://scottaaronson.com/blog/?p=87#comment-2150*

> "Revolution has already been televised. Revolution has been *merchandised*. Revolution is a commodity, a packaged lifestyle, available at your local mall. $19.95 gets you the black mask, the spray can, the "Crush the Fascists" protest sign, and access to your blog where you can write about the police brutality you suffered when you chained yourself to a fire hydrant. Capitalism has learned how to sell anti-capitalism."

Many in Silicon Valley have observed that the vast majority of venture capitalists at any given time are all chasing the same Revolutionary Innovation, and it's the Revolutionary Innovation that IPO'd six months ago. This is an *especially* crushing observation in venture capital, because there's a direct economic motive to not follow the herd—either someone else is also developing the product, or someone else is bidding too much for the startup. Steve Jurvetson once told me that at Draper Fisher Jurvetson, only two partners need to agree in order to fund any startup up to $1.5 million. And if *all* the partners agree that something sounds like a good idea, they won't do it. If only grant committees were this sane.

The problem with originality is that you actually have to *think* in order to attain it, instead of letting your brain complete the pattern[5]. There is no conveniently labeled "Outside the Box" to which you can immediately run off. There's an almost Zen-like quality to it—like the way you can't teach *satori* in words because *satori* is the experience of words failing you. The more you try to follow the Zen Master's instructions in words, the further you are from attaining an empty mind.

There is a reason, I think, why people do not attain novelty by striving for it. Properties like truth or good design are independent of novelty: 2 + 2 = 4, yes, really, even though this is what everyone else thinks too. People who strive to discover

4. *http://journalism.berkeley.edu/projects/biplog/archive/000748.html*
5. Page 297, 'Cached Thoughts'.

truth or to invent good designs, may in the course of time attain creativity. Not every change is an improvement, but every improvement is a change.

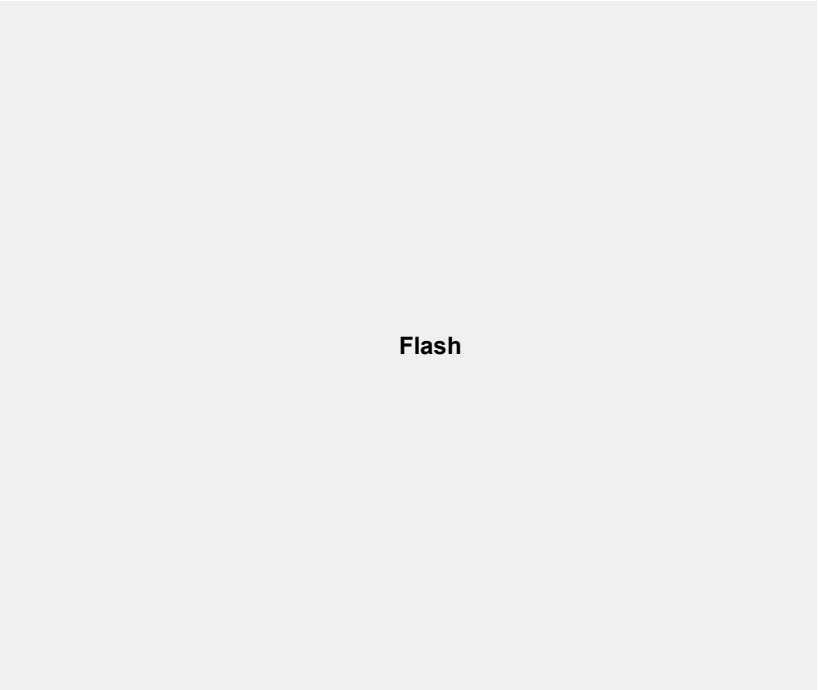
Every improvement is a change, but not every change is an improvement. The one who says, "I want to build an original mousetrap!", and not, "I want to build an optimal mousetrap!", nearly always wishes to be *perceived* as original. "Originality" in this sense is inherently social, because it can only be determined by comparison to other people. So their brain simply completes the standard pattern for what is perceived as "original", and their friends nod in agreement and say it is subversive.

Business books always tell you, for your convenience, where your cheese has been moved to. Otherwise the readers would be left around saying, "Where is this 'Outside the Box' I'm supposed to go?"

Actually thinking, like satori, is a wordless act of mind.

The eminent philosophers of Monty Python[6] said it best of all:

6. *http://www.youtube.com/watch?v=LQqq3e03EBQ*


Flash

6. Original Seeing[1]

Followup to: Cached Thoughts[2], The Virtue of Narrowness[3]

Since Robert Pirsig put this very well, I'll just copy down what he said. I don't know if this story is based on reality or not, but either way, it's true.

> He'd been having trouble with students who had nothing to say. At first he thought it was laziness but later it became apparent that it wasn't. They just couldn't think of anything to say.
>
> One of them, a girl with strong-lensed glasses, wanted to write a five-hundred word essay about the United States. He was used to the sinking feeling that comes from statements like this, and suggested without disparagement that she narrow it down to just Bozeman.
>
> When the paper came due she didn't have it and was quite upset. She had tried and tried but she just couldn't think of anything to say.
>
> It just stumped him. Now *he* couldn't think of anything to say. A silence occurred, and then a peculiar answer: "Narrow it down to the *main street* of Bozeman." It was a stroke of insight.
>
> She nodded dutifully and went out. But just before her next class she came back in *real* distress, tears this time, distress that had obviously been there for a long time. She still couldn't think of anything to say, and couldn't understand why, if she couldn't think of anything about *all* of Bozeman, she should be able to think of something about just one street.

1. *http://lesswrong.com/lw/k7/original_seeing/*
2. Page 297, 'Cached Thoughts'.
3. Page 58, 'The Virtue of Narrowness'.

He was furious. "You're not *looking!*" he said. A memory came back of his own dismissal from the University for having *too much* to say. For every fact there is an *infinity* of hypotheses. The more you *look* the more you *see*. She really wasn't looking and yet somehow didn't understand this.

He told her angrily, "Narrow it down to the *front* of *one* building on the main street of Bozeman. The Opera House. Start with the upper left-hand brick."

Her eyes, behind the thick-lensed glasses, opened wide.

She came in the next class with a puzzled look and handed him a five-thousand-word essay on the front of the Opera House on the main street of Bozeman, Montana. "I sat in the hamburger stand across the street," she said, "and started writing about the first brick, and the second brick, and then by the third brick it all started to come and I couldn't stop. They thought I was crazy, and they kept kidding me, but here it all is. I don't understand it."

Neither did he, but on long walks through the streets of town he thought about it and concluded she was evidently stopped with the same kind of blockage that had paralyzed him on his first day of teaching. She was blocked because she was trying to repeat, in her writing, things she had already heard, just as on the first day he had tried to repeat things he had already decided to say. She couldn't think of anything to write about Bozeman because she couldn't recall anything she had heard worth repeating. She was strangely unaware that she could look and see freshly for herself, as she wrote, without primary regard for what had been said before. The narrowing down to one brick destroyed the blockage because it was so obvious she *had* to do some original and direct seeing.

—Robert M. Pirsig, *Zen and the Art of Motorcycle Maintenance*

7. The Logical Fallacy of Generalization from Fictional Evidence[1]

When I try to introduce the subject of advanced AI, what's the first thing I hear, more than half the time?

"Oh, you mean like the Terminator movies / the Matrix / Asimov's robots!"

And I reply, "Well, no, not exactly. I try to avoid the logical fallacy of generalizing from fictional evidence."

Some people get it right away, and laugh. Others defend their use of the example, disagreeing that it's a fallacy.

What's wrong with using movies or novels as starting points for the discussion? No one's claiming that it's *true,* after all. Where is the lie, where is the rationalist sin? Science fiction represents the author's attempt to visualize the future; why not take advantage of the thinking that's already been done on our behalf, instead of starting over?

Not every misstep in the precise dance of rationality consists of outright belief in a falsehood; there are subtler ways to go wrong.

First, let us dispose of the notion that science fiction represents a full-fledged rational attempt to forecast the future. Even the most diligent science fiction writers are, first and foremost, storytellers; the requirements of storytelling are not the same as the requirements of forecasting. As Nick Bostrom points out[2]:

> "When was the last time you saw a movie about humankind suddenly going extinct (without warning and without being replaced by some other civilization)? While this scenario may be much more probable than a scenario in which human heroes

1. *http://lesswrong.com/lw/k9/the_logical_fallacy_of_generalization_from/*
2. *http://www.nickbostrom.com/existential/risks.html*

> successfully repel an invasion of monsters or robot warriors, it wouldn't be much fun to watch."

So there are specific distortions[3] in fiction. But trying to correct for these specific distortions is not enough. A story is *never* a rational attempt at analysis, not even with the most diligent science fiction writers, because stories don't use probability distributions. I illustrate as follows:

> Bob Merkelthud slid cautiously through the door of the alien spacecraft, glancing right and then left (or left and then right) to see whether any of the dreaded Space Monsters yet remained. At his side was the only weapon that had been found effective against the Space Monsters, a Space Sword forged of pure titanium with 30% probability, an ordinary iron crowbar with 20% probability, and a shimmering black discus found in the smoking ruins of Stonehenge with 45% probability, the remaining 5% being distributed over too many minor outcomes to list here.
>
> Merklethud (though there's a significant chance that Susan Wifflefoofer was there instead) took two steps forward or one step back, when a vast roar split the silence of the black airlock! Or the quiet background hum of the white airlock! Although Amfer and Woofi (1997) argue that Merklethud is devoured at this point, Spacklebackle (2003) points out that—

Characters can be ignorant, but the *author* can't say the three magic words "I don't know." The protagonist must thread a single line through the future, full of the details[4] that lend

3. *http://www.overcomingbias.com/2006/12/biases_of_scien.html*
4. *http://lesswrong.com/lw/jk/burdensome_details/*

flesh to the story, from Wifflefoofer's appropriately futuristic attitudes toward feminism, down to the color of her earrings.

Then all these burdensome details[5] and questionable assumptions are wrapped up and given a short label[6], creating the illusion that they are a single package[7].

On problems with large answer spaces, the greatest difficulty is not *verifying* the correct answer but simply locating it in answer space[8] to begin with. If someone starts out by asking whether or not AIs are gonna put us into capsules like in "The Matrix", they're jumping to a 100-bit proposition, without a corresponding 98 bits of evidence to locate it in the answer space as a possibility worthy of explicit consideration. It would only take a handful more evidence after the first 98 bits to promote that possibility to near-certainty, which tells you something about where nearly all the work gets done.

The "preliminary" step of locating possibilities worthy of explicit consideration includes steps like: Weighing what you know and don't know, what you can and can't predict, making a deliberate effort to avoid absurdity bias[9] and widen confidence intervals[10], pondering which questions are the important ones, trying to adjust for possible Black Swans and think of (formerly) unknown unknowns. Jumping to "*The Matrix:* Yes or No?" skips over all of this[11].

Any professional negotiator knows that to control the terms of a debate is very nearly to control the outcome of the debate. If you start out by thinking of *The Matrix,* it brings to mind marching robot armies defeating humans after a long struggle—not a superintelligence snapping nanotechnological fin-

5. *http://lesswrong.com/lw/jk/burdensome_details/*
6. Page 29, 'Occam's Razor'.
7. *http://en.wikipedia.org/wiki/Package-deal_fallacy*
8. *http://wiki.lesswrong.com/wiki/Locate_the_hypothesis*
9. *http://lesswrong.com/lw/j1/stranger_than_history/*
10. *http://lesswrong.com/lw/j6/why_is_the_future_so_absurd/*
11. *http://wiki.lesswrong.com/wiki/Privileging_the_hypothesis*

gers. It focuses on an "Us vs. Them" struggle, directing attention to questions like "Who will win?" and "Who should win?" and "Will AIs really be like that?" It creates a general atmosphere of entertainment, of "What is your amazing vision of the future?"

Lost to the echoing emptiness are: considerations of more than one possible mind design that an "Artificial Intelligence" could implement; the future's dependence on initial conditions; the power[12] of smarter-than-human intelligence and the argument for its unpredictability[13]; people taking the whole matter seriously and trying to do something about it.

If some insidious corrupter of debates decided that *their* preferred outcome would be best served by forcing discussants to start out by refuting *Terminator,* they would have done well in skewing the frame. Debating gun control, the NRA spokesperson does not wish to introduced as a "shooting freak", the anti-gun opponent does not wish to be introduced as a "victim disarmament advocate". Why should you allow the same order of frame-skewing by Hollywood scriptwriters, even accidentally?

Journalists don't tell me, "The future will be like *2001*". But they ask, "Will the future be like *2001*, or will it be like *A.I.*?" This is just as huge a framing issue as asking "Should we cut benefits for disabled veterans, or raise taxes on the rich?"

In the ancestral environment, there were no moving pictures; what you saw with your own eyes was true. A momentary glimpse of a single word can prime[14] us and make compatible thoughts more available[15], with demonstrated strong influence on probability estimates. How much havoc do you think a two-hour movie can wreak on your judgment? It will be hard enough to undo the damage by deliberate concentration—why

12. *http://intelligence.org/blog/2007/07/10/the-power-of-intelligence/*
13. *http://intelligence.org/blog/2007/09/30/three-major-singularity-schools/*
14. Page 291, 'Priming and Contamination'.
15. *http://lesswrong.com/lw/j5/availability/*

invite the vampire into your house? In Chess or Go, every wasted move is a loss; in rationality, any non-evidential influence is (on average) entropic.

Do movie-viewers succeed in unbelieving[16] what they see? So far as I can tell, few movie viewers act as if they have *directly* observed Earth's future. People who watched the *Terminator* movies didn't hide in fallout shelters on August 29, 1997. But those who commit the fallacy seem to act as if they had seen the movie events occurring on *some other* planet; not Earth, but somewhere similar to Earth.

You say, "Suppose we build a very smart AI," and they say, "But didn't that lead to nuclear war in *The Terminator?*" As far as I can tell, it's identical reasoning, down to the tone of voice, of someone who might say: "But didn't that lead to nuclear war on Alpha Centauri?" or "Didn't that lead to the fall of the Italian city-state of Piccolo in the fourteenth century?" The movie is not believed, but it is available[17]. It is treated, not as a prophecy, but as an illustrative historical case. Will history repeat itself? Who knows?

In a recent Singularity discussion, someone mentioned that Vinge didn't seem to think that brain-computer interfaces would increase intelligence much, and cited *Marooned in Realtime* and Tunç Blumenthal, who was the most advanced traveller but didn't seem all that powerful. I replied indignantly, "But Tunç lost most of his hardware! He was crippled!" And then I did a mental double-take and thought to myself: What the *hell* am I saying.

Does the issue not have to be argued in its own right, regardless of how Vinge depicted his characters? Tunç Blumenthal is not "crippled", he's *unreal.* I could say "Vinge chose to depict Tunç as crippled, for reasons that may or may not have had anything to do with his personal best forecast," and that would give his authorial choice an appropriate weight of evidence. I

16. Page 294, 'Do We Believe Everything We're Told?'.
17. *http://lesswrong.com/lw/j5/availability/*

cannot say "Tunç was crippled." There is no *was* of Tunç Blumenthal.

I deliberately left in a mistake I made, in my first draft of the top of this post: "Others defend their use of the *example,* disagreeing that it's a fallacy." But the Matrix is *not* an example!

A neighboring flaw is the logical fallacy of arguing from imaginary evidence: "Well, if you *did* go to the end of the rainbow, you *would* find a pot of gold—which just proves my point!" (Updating on evidence predicted, but not observed, is the mathematical mirror image of hindsight bias[18].)

The brain has many mechanisms for generalizing from observation, not just the availability heuristic. You see three zebras, you form the category "zebra", and this category embodies an automatic perceptual inference. Horse-shaped creatures with white and black stripes are classified as "Zebras", therefore they are fast and good to eat; they are expected to be similar to other zebras observed.

So people see (moving pictures of) three Borg, their brain automatically creates the category "Borg", and they infer automatically that humans with brain-computer interfaces are of class "Borg" and will be similar to other Borg observed: cold, uncompassionate, dressing in black leather, walking with heavy mechanical steps. Journalists don't believe that the future *will* contain Borg—they don't believe *Star Trek* is a prophecy. But when someone talks about brain-computer interfaces, they think, "Will the future contain Borg?" Not, "How do I know computer-assisted telepathy makes people less nice?" Not, "I've never seen a Borg and never has anyone else." Not, "I'm forming a racial stereotype based on *literally* zero evidence."

As George Orwell said[19] of cliches:

> "What is above all needed is to let the meaning
> choose the word, and not the other way around...

18. Page 87, 'Fake Causality'.
19. Page 180, 'Rationality and the English Language'.

> When you think of something abstract you are more inclined to use words from the start, and unless you make a conscious effort to prevent it, the existing dialect will come rushing in and do the job for you, at the expense of blurring or even changing your meaning."

Yet in my estimation, the *most* damaging aspect of using other authors' imaginations is that it stops people from using their own. As Robert Pirsig said[20]:

> "She was blocked because she was trying to repeat, in her writing, things she had already heard, just as on the first day he had tried to repeat things he had already decided to say. She couldn't think of anything to write about Bozeman because she couldn't recall anything she had heard worth repeating. She was strangely unaware that she could look and see freshly for herself, as she wrote, without primary regard for what had been said before."

Remembered fictions rush in and do your thinking for you; they substitute for *seeing*—the deadliest convenience of all.

Viewpoints taken here are further supported in: Anchoring[21], Contamination[22], Availability[23], Cached Thoughts[24], Do We Believe *Everything* We're Told?[25], Einstein's Arrogance[26], Burdensome details[27]

20. Page 304, 'Original Seeing'.
21. Page 289, 'Anchoring and Adjustment'.
22. Page 291, 'Priming and Contamination'.
23. *http://lesswrong.com/lw/j5/availability/*
24. Page 297, 'Cached Thoughts'.
25. Page 294, 'Do We Believe Everything We're Told?'.
26. *http://lesswrong.com/lw/jo/einsteins_arrogance/*
27. *http://lesswrong.com/lw/jk/burdensome_details/*

8. How to Seem (and Be) Deep[1]

I recently attended a discussion group whose topic, at that session, was Death. It brought out deep emotions. I think that of all the Silicon Valley lunches I've ever attended, this one was the most honest; people talked about the death of family, the death of friends, what they thought about their own deaths. People really listened to each other. I wish I knew how to reproduce those conditions reliably.

I was the only transhumanist present, and I was extremely careful not to be obnoxious about it. ("A fanatic is someone who can't change his mind and won't change the subject." I endeavor to at least be capable of changing the subject.) Unsurprisingly[2], people talked about the meaning that death gives to life, or how death is truly a blessing in disguise. But I did, very cautiously, explain that transhumanists are generally positive on life but thumbs down on death[3].

Afterward, several people came up to me and told me I was very "deep". Well, yes, I am, but this got me thinking about what makes people *seem* deep.

At one point in the discussion, a woman said that thinking about death led her to be nice to people because, who knows, she might not see them again. "When I have a nice thing to say about someone," she said, "now I say it to them right away, instead of waiting."

"That is a beautiful thought," I said, "and even if someday the threat of death is lifted from you, I hope you will keep on doing it—"

Afterward, this woman was one of the people who told me I was deep.

At another point in the discussion, a man spoke of some benefit X of death, I don't recall exactly what. And I said: "You

1. *http://lesswrong.com/lw/k8/how_to_seem_and_be_deep/*
2. Page 304, 'Original Seeing'.
3. *http://intelligence.org/blog/2007/06/16/transhumanism-as-simplified-humanism/*

know, given human nature, if people got hit on the head by a baseball bat every week, pretty soon they would invent reasons why getting hit on the head with a baseball bat was a good thing. But if you took someone who wasn't being hit on the head with a baseball bat, and you asked them if they wanted it, they would say no. I think that if you took someone who was immortal, and asked them if they wanted to die for benefit X, they would say no."

Afterward, this man told me I was deep.

Correlation is not causality. Maybe I was just speaking in a deep voice that day, and so sounded wise.

But my suspicion is that I came across as "deep" because I coherently violated the cached pattern[4] for "deep wisdom" in a way that made immediate sense.

There's a stereotype of Deep Wisdom. Death: complete the pattern: "Death gives meaning to life." Everyone knows this standard Deeply Wise response. And so it takes on some of the characteristics of an applause light. If you say it, people may nod along, because the brain completes the pattern and they know they're supposed to nod. They may even say "What deep wisdom!", perhaps in the hope of being thought deep themselves. But they will not be *surprised;* they will not have heard anything outside the box[5]*;* they will not have heard anything they could not have thought of for themselves. One might call it belief in wisdom[6]—the thought is labeled "deeply wise", and it's the completed standard pattern for "deep wisdom", but it carries no experience of insight.

People who *try to seem* Deeply Wise often end up seeming hollow, echoing[7] as it were, because they're trying to seem Deeply Wise instead of optimizing[8].

4. Page 297, 'Cached Thoughts'.
5. Page 300, 'The "Outside the Box" Box'.
6. Page 43, 'Belief in Belief'.
7. Page 304, 'Original Seeing'.
8. Page 300, 'The "Outside the Box" Box'.

How much thinking did I need to do, in the course of seeming deep? Human brains only run at 100Hz and I responded in realtime, so most of the work must have been precomputed. The part I experienced as effortful was picking a response understandable in one inferential step and then phrasing it for maximum impact.

Philosophically, nearly all of my work was already done. Complete the pattern: Existing condition X is really justified because it has benefit Y: "Naturalistic fallacy?" / "Status quo bias?" / "Could we get Y without X?" / "If we had never even heard of X before, would we voluntarily take it on to get Y?" I think it's fair to say that I execute these thought-patterns at around the same level of automaticity as I breathe. After all, most of human thought has to be cache lookups if the brain is to work at all.

And I already held to the developed philosophy of transhumanism[9]. Transhumanism also has cached thoughts about death. Death: complete the pattern: "Death is a pointless tragedy which people rationalize." This was a nonstandard cache, one with which my listeners were unfamiliar. I had several opportunities to use nonstandard cache, and because they were all part of the developed philosophy of transhumanism, they all visibly belonged to the same theme. This made me seem *coherent,* as well as original.

I suspect this is one reason Eastern philosophy seems deep to Westerners—it has nonstandard but coherent cache for Deep Wisdom. Symmetrically, in works of Japanese fiction, one sometimes finds Christians[10] depicted as repositories of deep wisdom and/or mystical secrets. (And sometimes not[11].)

If I recall correctly an economist once remarked that popular audiences are so unfamiliar with standard economics that, when he was called upon to make a television appearance, he

9. *http://intelligence.org/blog/2007/06/16/transhumanism-as-simplified-humanism/*
10. *http://tvtropes.org/pmwiki/pmwiki.php/Main/JesusTaboo*
11. *http://tvtropes.org/pmwiki/pmwiki.php/Main/CreepyCoolCrosses*

just needed to repeat back Econ 101 in order to sound like a brilliantly original thinker.

Also crucial was that my listeners could see *immediately* that my reply made sense. They might or might not have agreed with the thought, but it was not a complete non-sequitur unto them. I know transhumanists who are unable to seem deep because they are unable to appreciate what their listener does not already know. If you want to sound deep, you can never say anything that is more than a single step of inferential distance away from your listener's current mental state. That's just the way it is.

To *seem* deep, study nonstandard philosophies. Seek out discussions on topics that will give you a chance to appear deep. Do your philosophical thinking in advance, so you can concentrate on explaining well. Above all, practice staying within the one-inferential-step bound.

To *be* deep, think for yourself about "wise" or important or emotionally fraught topics. Thinking for yourself isn't the same as coming up with an unusual answer[12]. It does mean seeing for yourself[13], rather than letting your brain complete the pattern[14]. If you don't stop at the first answer[15], and cast out replies that seem vaguely unsatisfactory[16], in time your thoughts will form a coherent whole, flowing from the single source of yourself, rather than being fragmentary repetitions[17] of other people's conclusions[18].

12. Page 300, 'The "Outside the Box" Box'.
13. Page 304, 'Original Seeing'.
14. Page 297, 'Cached Thoughts'.
15. Page 426, 'The Third Alternative'.
16. Page 62, 'Your Strength as a Rationalist'.
17. Page 304, 'Original Seeing'.
18. Page 297, 'Cached Thoughts'.

9. We Change Our Minds Less Often Than We Think[1]

> "Over the past few years, we have discreetly approached colleagues faced with a choice between job offers, and asked them to estimate the probability that they will choose one job over another. The average confidence in the predicted choice was a modest 66%, but only 1 of the 24 respondents chose the option to which he or she initially assigned a lower probability, yielding an overall accuracy rate of 96%."
>
> —Dale Griffin and Amos Tversky, "The Weighing of Evidence and the Determinants of Confidence." (*Cognitive Psychology,* 24, pp. 411-435.)

When I first read the words above—on August 1st, 2003, at around 3 o'clock in the afternoon—it changed the way I thought. I realized that *once I could guess what my answer would be*—once I could assign a higher probability to deciding one way than other—then I had, in all probability, already decided. We change our minds less often than we think. And most of the time we become able to guess what our answer will be within half a second of hearing the question.

How swiftly that unnoticed moment passes, when we can't yet guess what our answer will be; the tiny window of opportunity for intelligence to act. In questions of choice, as in questions of fact.

The principle of the bottom line[2] is that only the actual causes of your beliefs determine your effectiveness as a rationalist. Once your belief is fixed, no amount of argument will alter the truth-value; once your decision is fixed, no amount of argument will alter the consequences.

---

1. *http://lesswrong.com/lw/jx/we_change_our_minds_less_often_than_we_think/*
2. Page 343, 'The Bottom Line'.

You might think that you could arrive at a belief, or a decision, by non-rational means, and then try to justify it, and if you found you couldn't justify it, reject it.

But we change our minds less often—*much* less often—than we think.

I'm sure that you can think of at least one occasion in your life when you've changed your mind. We all can. How about all the occasions in your life when you didn't change your mind? Are you they as available[3], in your heuristic estimate of your competence[4]?

Between hindsight bias[5], fake causality[6], positive bias[7], anchoring[8]/priming, et cetera et cetera, and above all the dreaded confirmation bias[9], once an idea gets into your head, it's probably going to stay there.

3. *http://lesswrong.com/lw/j5/availability/*
4. *http://en.wikipedia.org/wiki/Lake_Wobegon_effect*
5. Page 71, 'Hindsight bias'.
6. Page 87, 'Fake Causality'.
7. Page 108, 'Positive Bias: Look Into the Dark'.
8. Page 289, 'Anchoring and Adjustment'.
9. Page 333, 'Knowing About Biases Can Hurt People'.

10. Hold Off On Proposing Solutions[1]

From pp. 55-56 of Robyn Dawes's *Rational Choice in an Uncertain World*. Bolding added.

> Norman R. F. Maier noted that when a group faces a problem, the natural tendency of its members is to propose possible solutions as they begin to discuss the problem. Consequently, the group interaction focuses on the merits and problems of the proposed solutions, people become emotionally attached to the ones they have suggested, and superior solutions are not suggested. Maier enacted an edict to enhance group problem solving: "**Do not propose solutions until the problem has been discussed as thoroughly as possible without suggesting any.**" It is easy to show that this edict works in contexts where there are objectively defined good solutions to problems.
>
> Maier devised the following "role playing" experiment to demonstrate his point. Three employees of differing ability work on an assembly line. They rotate among three jobs that require different levels of ability, because the most able—who is also the most dominant—is strongly motivated to avoid boredom. In contrast, the least able worker, aware that he does not perform the more difficult jobs as well as the other two, has agreed to rotation because of the dominance of his able co-worker. An "efficiency expert" notes that if the most able employee were given the most difficult task and the least able the least difficult, productivity could be improved by 20%, and the expert recommends that

1. *http://lesswrong.com/lw/ka/hold_off_on_proposing_solutions/*

> the employees stop rotating. The three employees and the a fourth person designated to play the role of foreman are asked to discuss the expert's recommendation. Some role-playing groups are given Maier's edict not to discuss solutions until having discussed the problem thoroughly, while others are not. Those who are not given the edict immediately begin to argue about the importance of productivity versus worker autonomy and the avoidance of boredom. Groups presented with the edict have a much higher probability of arriving at the solution that the two more able workers rotate, while the least able one sticks to the least demanding job—a solution that yields a 19% increase in productivity.
>
> I have often used this edict with groups I have led—**particularly when they face a very tough problem, which is when group members are most apt to propose solutions immediately.** While I have no objective criterion on which to judge the quality of the problem solving of the groups, Maier's edict appears to foster better solutions to problems.

This is so true it's not even funny. And it gets worse and worse the tougher the problem becomes. Take Artificial Intelligence, for example. A surprising number of people I meet seem to know exactly how to build an Artificial General Intelligence, without, say, knowing how to build an optical character recognizer or a collaborative filtering system (much easier problems). And as for building an AI with a positive impact on the world—a Friendly AI[2], loosely speaking—why, *that* problem is so incredibly difficult that an actual *majority* resolve the whole issue within 15 seconds. *Give* me a *break*.

2. *http://intelligence.org/AIRisk.pdf*

(**Added**: This problem is by no means unique to AI. Physicists encounter plenty of nonphysicists with their own theories of physics, economists get to hear lots of amazing new theories of economics. If you're an evolutionary biologist, anyone you meet can instantly solve any open problem in your field, usually by postulating group selection. Et cetera.)

Maier's advice echoes the principle of the bottom line[3], that the effectiveness of our decisions is determined only by whatever evidence and processing we did in first arriving at our decisions—after you write the bottom line, it is too late to write more reasons[4] above. If you make your decision very early on, it will, in fact, be based on very little thought, no matter how many amazing arguments you come up with afterward.

And consider furthermore that We Change Our Minds Less Often Than We Think[5]: 24 people assigned an average 66% probability to the future choice thought more probable, but only 1 in 24 actually chose the option thought less probable. **Once you can guess what your answer will be, you have probably already decided.** If you can guess your answer half a second after hearing the question, then you have half a second in which to be intelligent. It's not a lot of time.

Traditional Rationality[6] emphasizes *falsification*— the ability to *relinquish* an initial opinion when confronted by clear evidence against it. But once an idea gets into your head, it will probably require way too much evidence to get it out again. Worse, we don't always have the luxury of overwhelming evidence.

I suspect that a more powerful (and more difficult) method is to *hold off on thinking of an answer*. To suspend, draw out, that tiny moment when we can't yet guess what our answer will be; thus giving our intelligence a longer time in which to act.

3. Page 343, 'The Bottom Line'.
4. Page 351, 'Rationalization'.
5. Page 318, 'We Change Our Minds Less Often Than We Think'.
6. Page 489, 'No One Can Exempt You From Rationality's Laws'.

Even half a minute would be an improvement over half a second.

11. On Expressing Your Concerns[1]

Followup to: Asch's Conformity Experiment[2]

The scary thing about Asch's conformity experiments[3] is that you can get many people to say black is white, if you put them in a room full of other people saying the same thing. The hopeful thing about Asch's conformity experiments is that a single dissenter tremendously drove down the rate of conformity, even if the dissenter was only giving a different wrong answer. And the *wearisome* thing is that dissent was not *learned* over the course of the experiment—when the single dissenter started siding with the group, rates of conformity rose back up.

Being a voice of dissent can bring real benefits to the group. But it also (famously) has a cost. And then you have to keep it up. Plus you could be wrong.

I recently had an interesting experience wherein I began discussing a project with two people who had previously done some planning on their own. I thought they were being too optimistic[4] and made a number of safety-margin-type suggestions for the project. Soon a fourth guy wandered by, who was providing one of the other two with a ride home, and began making suggestions. At this point I had a sudden insight about how groups become overconfident, because whenever I raised a possible problem, the fourth guy would say, "Don't worry, I'm sure we can handle it!" or something similarly reassuring.

An individual, working alone, will have natural doubts. They will think to themselves, "Can I really do XYZ?", because there's nothing impolite about doubting your *own* competence. But when two unconfident people form a group, it is polite to say nice and reassuring things, and impolite to question the other person's competence. Together they become more opti-

1. *http://lesswrong.com/lw/ma/on_expressing_your_concerns/*
2. Page 267, 'Asch's Conformity Experiment'.
3. Page 267, 'Asch's Conformity Experiment'.
4. *http://lesswrong.com/lw/jg/planning_fallacy/*

mistic than either would be on their own, each one's doubts quelled by the other's seemingly confident reassurance, not realizing that the other person initially had the same inner doubts.

The most fearsome possibility raised by Asch's experiments on conformity is the specter of everyone agreeing with the group, swayed by the confident voices of others, careful not to let their own doubts show—not realizing that others are suppressing similar worries. This is known as "pluralistic ignorance".

Robin Hanson and I have a long-running debate over when, exactly, aspiring rationalists should dare to disagree. I tend toward the widely held position that you have no real choice but to form your own opinions. Robin Hanson advocates a more iconoclastic position, that *you*—not just other people—should consider that others may be wiser. Regardless of our various disputes, we both agree that Aumann's Agreement Theorem extends to imply that common knowledge of a factual[5] disagreement shows *someone* must be irrational[6]. Despite the funny looks we've gotten, we're sticking to our guns about modesty: Forget what everyone tells you about individualism, you *should* pay attention to what other people think.

Ahem. The point is that, for rationalists, disagreeing with the group is serious business. You can't wave it off with "Everyone is entitled to their own opinion.[7]"

I think the most important lesson to take away from Asch's experiments is to distinguish "expressing concern" from "disagreement". Raising a point that others haven't voiced is not a promise to disagree with the group at the end of its discussion.

The ideal Bayesian's process of convergence involves sharing evidence that is unpredictable to the listener. The Aumann agreement result holds only for *common knowledge,* where you

5. Page 463, 'Feeling Rational'.
6. *http://lesswrong.com/lw/gr/the_modesty_argument/*
7. *http://www.overcomingbias.com/2006/12/you_are_never_e.html*

know, I know, you know I know, etc. Hanson's post or paper on "We Can't Foresee to Disagree[8]" provides a picture of how strange it would look to watch ideal rationalists converging on a probability estimate; it doesn't look anything like two bargainers in a marketplace converging on a price.

Unfortunately, there's not much difference *socially* between "expressing concerns" and "disagreement". A group of rationalists might agree to pretend there's a difference, but it's not how human beings are really wired. Once you speak out, you've committed a socially irrevocable act; you've become the nail sticking up, the discord in the comfortable group harmony, and you can't undo that. Anyone insulted by a concern you expressed about their competence to successfully complete task XYZ, will probably hold just as much of a grudge afterward if you say "No problem, I'll go along with the group" at the end.

Asch's experiment shows that the power of dissent to inspire others is real. Asch's experiment shows that the power of conformity is real. If everyone refrains from voicing their private doubts, that will indeed lead groups into madness. But history abounds with lessons on the price of being the first, or even the second, to say that the Emperor has no clothes. Nor are people hardwired to distinguish "expressing a concern" from "disagreement even with common knowledge"; this distinction is a rationalist's artifice. If you read the more cynical brand of self-help books (e.g. Machiavelli's *The Prince*) they will advise you to mask your nonconformity entirely, *not* voice your concerns first and then agree at the end. If you perform the group service of being the one who gives voice to the obvious problems, don't expect the group to thank you for it.

These are the costs and the benefits of dissenting—whether you "disagree" or just "express concern"—and the decision is up to you.

8. *http://www.overcomingbias.com/2007/01/we_cant_foresee.html*

12. The Genetic Fallacy[1]

In lists[2] of[3] logical[4] fallacies[5], you will find included "the genetic fallacy"—the fallacy attacking a belief, based on someone's causes for believing it.

This is, at first sight, a very strange idea—if the causes of a belief do not determine its systematic reliability, what does? If Deep Blue advises us of a chess move, we trust it based on our understanding of the *code* that searches the game tree, being unable to evaluate the actual game tree ourselves. What could license any probability assignment as "rational", except that it was produced by some systematically reliable process?

Articles on the genetic fallacy will tell you that genetic reasoning is not always a fallacy—that the origin of evidence *can* be relevant to its evaluation, as in the case of a trusted expert. But other times, say[6] the articles, it *is* a fallacy; the chemist Kekulé first saw the ring structure of benzene in a dream, but this doesn't mean we can never trust this belief.

So sometimes the genetic fallacy is a fallacy, and sometimes it's not?

The genetic fallacy is formally a fallacy, because the *original cause* of a belief is not the same as its *current justificational status,* the sum of all the support and antisupport *currently* known.

Yet we change our minds less often than we think[7]. Genetic accusations have a force among humans that they would not have among ideal Bayesians.

---

1. *http://lesswrong.com/lw/s3/the_genetic_fallacy/*
2. *http://www.nizkor.org/features/fallacies/genetic-fallacy.html*
3. *http://en.wikipedia.org/wiki/List_of_fallacies*
4. *http://atheism.about.com/library/FAQs/skepticism/blfaq_fall_genetic.htm*
5. *http://www.fallacyfiles.org/genefall.html*
6. *http://www.fallacyfiles.org/genefall.html*
7. Page 318, 'We Change Our Minds Less Often Than We Think'.

Clearing your mind is a *powerful heuristic* when you're faced with new suspicion that many of your ideas may have come from a flawed source.

Once an idea gets into our heads, it's not always easy for evidence to root it out. Consider all the people out there who grew up believing in the Bible; later came to reject (on a deliberate level) the idea that the Bible was written by the hand of God; and who nonetheless think that the Bible contains indispensable ethical wisdom[8]. They have failed to clear their minds; they could do significantly better by doubting anything the Bible said *because the Bible said it*.

At the same time, they would have to bear firmly in mind the principle that reversed stupidity is not intelligence[9]; the goal is to genuinely shake your mind loose and do independent thinking, not to negate the Bible and let that be your algorithm.

Once an idea gets into your head, you tend to find support for it everywhere you look—and so when the original source is suddenly cast into suspicion, you would be very wise indeed to suspect all the leaves that originally grew on that branch...

If you can! It's not easy to clear your mind. It takes a convulsive effort to *actually reconsider*, instead of letting your mind fall into the pattern of rehearsing[10] cached[11] arguments. "It ain't a true crisis of faith unless things could just as easily go either way," said Thor Shenkel.

You should be *extremely suspicious* if you have many ideas suggested by a source that you now know to be untrustworthy, but by golly, it seems that all the ideas still ended up being right—the Bible being the obvious archetypal example.

On the other hand... there's such a thing as sufficiently clear-cut evidence, that it no longer significantly matters where the idea originally came from. Accumulating that kind of clear-

8. *http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/*
9. Page 168, 'Reversed Stupidity Is Not Intelligence'.
10. Page 340, 'One Argument Against An Army'.
11. Page 297, 'Cached Thoughts'.

cut evidence is what Science[12] is all about. It doesn't matter any more that Kekulé first saw the ring structure of benzene in a dream—it wouldn't matter if we'd found the hypothesis to test[13] by generating random computer images, or from a spiritualist revealed as a fraud, or even from the Bible. The ring structure of benzene is pinned down by enough experimental evidence to make the source of the suggestion irrelevant.

In the absence of such clear-cut evidence, then you do need to pay attention to the original sources of ideas—to give experts more credence than layfolk, if their field has earned respect—to suspect ideas you originally got from suspicious sources—to distrust those whose motives are untrustworthy, *if* they cannot present arguments independent of their own authority.

The genetic fallacy is a *fallacy* when there exist justifications *beyond* the genetic fact asserted, but the genetic accusation is presented as if it settled the issue.

Some good rules of thumb (for humans):

- Be suspicious of genetic accusations against beliefs that you dislike, especially if the proponent claims justifications beyond the simple authority of a speaker. "Flight is a religious idea, so the Wright Brothers must be liars" is one of the classically given examples.
- By the same token, don't think you can get good information about a technical issue just by sagely psychoanalyzing the personalities involved and their flawed motives. If technical arguments exist, they get priority.
- When new suspicion is cast on one of your fundamental sources, you really *should* doubt all the branches and leaves that grew from that root. You are not licensed to reject them outright as conclusions, because reversed stupidity is not intelligence, but...

12. *http://lesswrong.com/lw/in/scientific_evidence_legal_evidence_rational/*
13. *http://lesswrong.com/lw/jo/einsteins_arrogance/*

- Be extremely suspicious if you find that you still believe the early suggestions of a source you later rejected.

Added: Hal Finney suggests[14] that we should call it "the genetic heuristic".

14. *http://www.overcomingbias.com/2008/07/genetic-fallacy.html#comment-121978890*

Against Rationalization

1. Knowing About Biases Can Hurt People[1]

Once upon a time I tried to tell my mother about the problem of expert calibration, saying: "So when an expert says they're 99% confident, it only happens about 70% of the time." Then there was a pause as, suddenly, I realized I was talking to my mother, and I hastily added: "Of course, you've got to make sure to apply that skepticism evenhandedly, including to yourself, rather than just using it to argue against anything you disagree with—"

And my mother said: "Are you kidding? This is great! I'm going to use it all the time!"

Taber and Lodge's Motivated skepticism in the evaluation of political beliefs[2] describes the confirmation of six predictions:

1. Prior attitude effect. Subjects who feel strongly about an issue—even when encouraged to be objective—will evaluate supportive arguments more favorably than contrary arguments.
2. Disconfirmation bias. Subjects will spend more time and cognitive resources denigrating contrary arguments than supportive arguments.
3. Confirmation bias. Subjects free to choose their information sources will seek out supportive rather than contrary sources.
4. **Attitude polarization. Exposing subjects to an apparently balanced set of pro and con arguments will exaggerate their initial polarization.**
5. Attitude strength effect. Subjects voicing stronger attitudes will be more prone to the above biases.
6. **Sophistication effect. Politically knowledgeable subjects, because they possess greater ammunition with which to counter-**

1. *http://lesswrong.com/lw/he/knowing_about_biases_can_hurt_people/*
2. *http://www.sunysb.edu/polsci/mlodge/lodgemotivated.pdf*

argue incongruent facts and arguments, will be more prone to the above biases.

If you're irrational to start with, having *more* knowledge can *hurt* you. For a true Bayesian, information would never have negative expected utility. But humans aren't perfect Bayes-wielders; if we're not careful, we can cut ourselves.

I've *seen* people severely messed up by their own knowledge of biases. They have more ammunition with which to argue against anything they don't like. And that problem—too much ready ammunition—is one of the primary ways that people with high mental agility end up stupid, in Stanovich's "dysrationalia" sense of stupidity.

You can think of people who fit this description, right? People with high g-factor who end up being *less* effective because they are too sophisticated as arguers? Do you think you'd be helping them—making them more effective rationalists—if you just told them about a list of classic biases?

I recall someone who learned about the calibration / overconfidence problem. Soon after he said: "Well, you can't trust experts; they're wrong so often as experiments have shown. So therefore, when I predict the future, I prefer to assume that things will continue historically as they have—" and went off into this whole complex, error-prone, highly questionable extrapolation. Somehow, when it came to trusting his own preferred conclusions, all those biases and fallacies seemed much less *salient*—leapt much less readily to mind—than when he needed to counter-argue someone else.

I told the one about the problem of disconfirmation bias and sophisticated argument, and lo and behold, the next time I said something he didn't like, he accused me of being a sophisticated arguer. He didn't try to point out any particular sophisticated argument, any particular flaw—just shook his head and sighed sadly over how I was apparently using my own intelligence to defeat itself. He had acquired yet another Fully General Counterargument.

Even the notion of a "sophisticated arguer" can be deadly, if it leaps all too readily to mind when you encounter a seemingly intelligent person who says something you don't like.

I endeavor to learn from my mistakes. The last time I gave a talk on heuristics and biases, I started out by introducing the general concept by way of the conjunction fallacy and representativeness heuristic. And then I moved on to confirmation bias, disconfirmation bias, sophisticated argument, motivated skepticism, and other attitude effects. I spent the next thirty minutes *hammering* on that theme, reintroducing it from as many different perspectives as I could.

I wanted to get my audience interested in the subject. Well, a simple description of conjunction fallacy and representativeness would suffice for that. But suppose they did get interested. Then what? The literature on bias is mostly cognitive psychology for cognitive psychology's sake. I had to give my audience their dire warnings during that one lecture, or they probably wouldn't hear them at all.

Whether I do it on paper, or in speech, I now try to never mention calibration and overconfidence unless I have first talked about disconfirmation bias, motivated skepticism, sophisticated arguers, and dysrationalia in the mentally agile. First, do no harm!

2. Update Yourself Incrementally[1]

Politics is the mind-killer[2]. Debate is war, arguments are soldiers[3]. There is the temptation to search for ways to interpret every possible experimental result[4] to confirm your theory, like securing a citadel against every possible line of attack. This you cannot do. It is mathematically impossible. For every expectation of evidence, there is an equal and opposite expectation of counterevidence.[5]

But it's okay if your cherished belief isn't *perfectly* defended. If the hypothesis is that the coin comes up heads 95% of the time, then one time in twenty you will see what looks like contrary evidence. This is okay. It's normal. It's even expected, so long as you've got nineteen supporting observations for every contrary one. A probabilistic model can take a hit or two[6], and still survive, so long as the hits don't *keep on* coming in.

Yet it is widely believed, especially in the court of public opinion, that a true theory can have *no* failures and a false theory *no* successes.

You find people holding up a single piece of what they conceive to be evidence, and claiming that their theory can 'explain' it, as though this were all the support that any theory needed. Apparently a false theory can have *no* supporting evidence; it is impossible for a false theory to fit even a single event. Thus, a single piece of confirming evidence is all that any theory needs.

It is only slightly less foolish to hold up a single piece of *probabilistic* counterevidence as disproof, as though it were impossible for a correct theory to have even a *slight* argument against it. But this is how humans have argued for ages and

1. *http://lesswrong.com/lw/ij/update_yourself_incrementally/*
2. Page 148, 'Politics is the Mind-Killer'.
3. Page 150, 'Policy Debates Should Not Appear One-Sided'.
4. Page 43, 'Belief in Belief'.
5. Page 68, 'Conservation of Expected Evidence'.
6. *http://lesswrong.com/lw/ig/i_defy_the_data/*

ages, trying to defeat all enemy arguments, while denying the enemy even a single shred of support. People want their debates to be one-sided; they are accustomed to a world in which their preferred theories have not one iota of antisupport. Thus, allowing a single item of probabilistic counterevidence would be the end of the world.

I just know someone in the audience out there is going to say, "But you *can't* concede even a single point if you want to win debates in the real world! If you concede that any counterarguments exist, the Enemy will harp on them over and over—you can't let the Enemy do that! You'll *lose!* What could be more viscerally terrifying than *that?*"

Whatever. Rationality is not for winning debates, it is for deciding which side to join. If you've already decided which side to argue for, the work of rationality is *done* within you, whether well or poorly. But how can you, yourself, decide which side to argue? If *choosing the wrong side* is viscerally terrifying, even just a little viscerally terrifying, you'd best integrate *all* the evidence.

Rationality is not a walk, but a dance. On each step in that dance your foot should come down in exactly the correct spot, neither to the left nor to the right. Shifting belief upward with each iota of confirming evidence. Shifting belief downward with each iota of contrary evidence. Yes, *down*. Even with a correct model, if it is not an exact model, you will sometimes need to revise your belief *down*.

If an iota or two of evidence happens to countersupport your belief, that's okay. It happens, sometimes, with probabilistic evidence for non-exact theories. (If an exact theory fails, you *are* in trouble!) Just shift your belief downward a little—the probability, the odds ratio, or even a nonverbal weight of credence in your mind. Just shift downward a little, and wait for more evidence[7]. If the theory is true, supporting evidence will

7. Page 68, 'Conservation of Expected Evidence'.

come in shortly, and the probability will climb again. If the theory is false, you don't really want it anyway.

The problem with using black-and-white, binary, qualitative reasoning is that any single observation either destroys the theory or it does not. When not even a single contrary observation is allowed, it creates cognitive dissonance and has to be argued away[8]. And this rules out incremental progress; it rules out correct integration of all the evidence. Reasoning probabilistically, we realize that on average, a correct theory will generate a greater weight of support than countersupport. And so you can, *without fear,* say to yourself: "This is gently contrary evidence, I will shift my belief downward". Yes, *down*. It does not destroy your cherished theory. That is qualitative reasoning; think quantitatively.

For every expectation of evidence, there is an equal and opposite expectation of counterevidence.[9] On every occasion, you must, on average, anticipate revising your beliefs downward as much as you anticipate revising them upward. If you think you already know what evidence will come in, then you must already be fairly sure of your theory—probability close to 1—which doesn't leave much room for the probability to go further upward. And however unlikely it seems that you will encounter disconfirming evidence, the resulting downward shift must be large enough to precisely balance the anticipated gain on the other side. The weighted mean of your expected posterior probability must equal your prior probability.

How silly is it, then, to be terrified[10] of revising your probability downward, if you're bothering to investigate a matter at all? On average, you must anticipate as much downward shift as upward shift from every individual observation.

It may perhaps happen that an iota of antisupport comes in again, and again and again, while new support is slow to

8. *http://lesswrong.com/lw/ig/i_defy_the_data/*
9. Page 68, 'Conservation of Expected Evidence'.
10. *http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/*

trickle in. You may find your belief drifting downward and further downward. Until, finally, you realize from which quarter the winds of evidence are blowing against you. In that moment of realization, there is no point in constructing excuses. In that moment of realization, you have *already relinquished* your cherished belief. Yay! Time to celebrate! Pop a champagne bottle or send out for pizza! You can't become stronger[11] by keeping the beliefs you started with, after all.

11. *http://lesswrong.com/lw/h8/tsuyoku_naritai_i_want_to_become_stronger/*

3. One Argument Against An Army[1]

Followup to: Update Yourself Incrementally[2]

Yesterday I talked about a style of reasoning in which not a single contrary argument is allowed, with the result that every non-supporting observation has to be argued away[3]. Today I suggest that when people encounter a contrary argument, they prevent themselves from downshifting their confidence by *rehearsing* already-known support.

Suppose the country of Freedonia is debating whether its neighbor, Sylvania, is responsible for a recent rash of meteor strikes on its cities. There are several pieces of evidence suggesting this: the meteors struck cities close to the Sylvanian border; there was unusual activity in the Sylvanian stock markets *before* the strikes; and the Sylvanian ambassador Trentino was heard muttering about "heavenly vengeance".

Someone comes to you and says: "I don't think Sylvania is responsible for the meteor strikes. They have trade with us of billions of dinars annually." "Well," you reply, "the meteors struck cities close to Sylvania, there was suspicious activity in their stock market, and their ambassador spoke of heavenly vengeance afterward." Since these three arguments outweigh the first, you *keep* your belief that Sylvania is responsible—you believe rather than disbelieve, qualitatively. Clearly, the balance of evidence weighs against Sylvania.

Then another comes to you and says: "I don't think Sylvania is responsible for the meteor strikes. Directing an asteroid strike is really hard. Sylvania doesn't even have a space program." You reply, "But the meteors struck cities close to Sylvania, and their investors knew it, and the ambassador came right out and admitted it!" Again, these three arguments out-

1. *http://lesswrong.com/lw/ik/one_argument_against_an_army/*
2. Page 336, 'Update Yourself Incrementally'.
3. Page 336, 'Update Yourself Incrementally'.

weigh the first (by three arguments against one argument), so you keep your belief that Sylvania is responsible.

Indeed, your convictions are *strengthened*. On two separate occasions now, you have evaluated the balance of evidence, and both times the balance was tilted against Sylvania by a ratio of 3-to-1.

You encounter further arguments by the pro-Sylvania traitors—again, and again, and a hundred times again—but each time the new argument is handily defeated by 3-to-1. And on every occasion, you feel yourself becoming more confident that Sylvania was indeed responsible, shifting your prior according to the felt balance of evidence.

The problem, of course, is that by *rehearsing* arguments you *already knew,* you are double-counting the evidence This would be a grave sin even if you double-counted *all* the evidence. (Imagine a scientist who does an experiment with 50 subjects and fails to obtain statistically significant results, so he counts all the data twice.)

But to selectively double-count *only some* evidence is sheer farce. I remember seeing a cartoon as a child, where a villain was dividing up loot using the following algorithm: "One for you, one for me. One for you, one-two for me. One for you, one-two-three for me."

As I emphasized yesterday,[4] even if a cherished belief is *true,* a rationalist may sometimes need to downshift the probability while integrating *all* the evidence. Yes, the balance of support may still favor your cherished belief. But you still have to shift the probability *down*—yes, *down*—from whatever it was before you heard the contrary evidence. It does no good to *rehearse* supporting arguments, because you have already taken those into account.

And yet it does appear to me that when people are confronted by a *new* counterargument, they search for a justification not to downshift their confidence, and of course they find sup-

4. Page 336, 'Update Yourself Incrementally'.

porting arguments they *already know*. I have to keep constant vigilance not to do this myself! It feels as natural as parrying a sword-strike with a handy shield.

With the right kind of wrong reasoning, a handful of support—or even a single argument—can stand off an army of contradictions.

4. The Bottom Line[1]

There are two sealed boxes up for auction, box A and box B. One and only one of these boxes contains a valuable diamond. There are all manner of signs and portents indicating whether a box contains a diamond; but I have no sign which I *know* to be perfectly reliable. There is a blue stamp on one box, for example, and I know that boxes which contain diamonds are more likely than empty boxes to show a blue stamp. Or one box has a shiny surface, and I have a suspicion—I am not sure—that no diamond-containing box is ever shiny.

Now suppose there is a clever arguer, holding a sheet of paper, and he says to the owners of box A and box B: "Bid for my services, and whoever wins my services, I shall argue that their box contains the diamond, so that the box will receive a higher price." So the box-owners bid, and box B's owner bids higher, winning the services of the clever arguer.

The clever arguer begins to organize his thoughts. First, he writes, "And *therefore,* box B contains the diamond!" at the bottom of his sheet of paper. Then, at the top of the paper, he writes, "Box B shows a blue stamp," and beneath it, "Box A is shiny", and then, "Box B is lighter than box A", and so on through many signs and portents; yet the clever arguer neglects all those signs which might argue in favor of box A. And then the clever arguer comes to me and recites from his sheet of paper: "Box B shows a blue stamp, and box A is shiny," and so on, until he reaches: "And *therefore,* box B contains the diamond."

But consider: At the moment when the clever arguer wrote down his conclusion, at the moment he put ink on his sheet of paper, the evidential entanglement[2] of that physical ink with the physical boxes became fixed.

It may help to visualize a collection of worlds—Everett branches or Tegmark duplicates[3]—within which there is some

1. *http://lesswrong.com/lw/js/the_bottom_line/*
2. Page 18, 'What is Evidence?'.

objective frequency at which box A or box B contains a diamond. There's likewise some objective frequency within the subset "worlds with a shiny box A" where box B contains the diamond; and some objective frequency in "worlds with shiny box A and blue-stamped box B" where box B contains the diamond.

The ink on paper is formed into odd shapes and curves, which look like this text: "And *therefore,* box B contains the diamond." If you happened to be a literate English speaker, you might become confused, and think that this shaped ink somehow *meant* that box B contained the diamond. Subjects instructed to say the color of printed pictures and shown the picture "green" often say "green" instead of "red". It helps to be illiterate, so that you are not confused by the shape of the ink.

To us, the true import of a thing is its entanglement with other things. Consider again the collection of worlds, Everett branches or Tegmark duplicates. At the moment when all clever arguers in all worlds put ink to the bottom line of their paper—let us suppose this is a single moment—it fixed the correlation of the ink with the boxes. The clever arguer writes in non-erasable pen; the ink will not change. The boxes will not change. Within the subset of worlds where the ink says "And therefore, box B contains the diamond," there is already some fixed percentage of worlds where box A contains the diamond. This will not change regardless of what is written in on the blank lines above.

So the evidential entanglement of the ink is fixed, and I leave to you to decide what it might be. Perhaps box owners who believe a better case can be made for them are more liable to hire advertisers; perhaps box owners who fear their own deficiencies bid higher. If the box owners do not themselves understand the signs and portents, then the ink will be completely unentangled with the boxes' contents, though it may tell you something about the owners' finances and bidding habits.

3. *http://arxiv.org/abs/astro-ph/0302131*

Now suppose another person present is genuinely curious, and she *first* writes down all the distinguishing signs of *both* boxes on a sheet of paper, and then applies her knowledge and the laws of probability and writes down at the bottom: "*Therefore,* I estimate an 85% probability that box B contains the diamond." Of what is this handwriting evidence? Examining the chain of cause and effect leading to this physical ink on physical paper, I find that the chain of causality wends its way through all the signs and portents of the boxes, and is dependent on these signs; for in worlds with different portents, a different probability is written at the bottom.

So the handwriting of the curious inquirer is entangled with the signs and portents and the contents of the boxes, whereas the handwriting of the clever arguer is evidence only of which owner paid the higher bid. There is a great difference in the indications of ink, though one who foolishly read aloud the ink-shapes might think the English words sounded similar.

Your effectiveness as a rationalist is determined by whichever algorithm actually writes the bottom line of your thoughts. If your car makes metallic squealing noises when you brake, and you aren't willing to face up to the financial cost of getting your brakes replaced, you can decide to look for reasons why your car might not need fixing. But the actual percentage of you that survive in Everett branches or Tegmark worlds—which we will take to describe your effectiveness as a rationalist—is determined by the algorithm that decided *which* conclusion you would seek arguments for. In this case, the real algorithm is "Never repair anything expensive." If this is a good algorithm, fine; if this is a bad algorithm, oh well. The arguments you write afterward, above the bottom line, will not change anything either way.

Addendum: This is intended as a caution for your own thinking, not a Fully General Counterargument against conclusions you don't like. For it is indeed a clever argument to say "My opponent is a clever arguer", if you are paying yourself to

retain whatever beliefs you had at the start. The world's cleverest arguer may point out that the sun is shining, and yet it is still probably daytime. See What Evidence Filtered Evidence?[4] for more on this topic.

4. Page 347, 'What Evidence Filtered Evidence?'.

5. What Evidence Filtered Evidence?[1]

Yesterday I discussed the dilemma of the clever arguer[2], hired to sell you a box that may or may not contain a diamond. The clever arguer points out to you that the box has a blue stamp, and it is a valid known fact that diamond-containing boxes are more likely than empty boxes to bear a blue stamp. What happens at this point, from a Bayesian perspective? Must you helplessly update your probabilities, as the clever arguer wishes?

If you can look at the box yourself, you can add up all the signs yourself. What if you can't look? What if the only evidence you have is the word of the clever arguer, who is legally constrained to make only true statements, but does not tell you everything he knows? Each statement that he makes is valid evidence—how could you *not* update your probabilities? Has it ceased to be true that, in such-and-such a proportion of Everett branches or Tegmark duplicates in which box B has a blue stamp, box B contains a diamond? According to Jaynes, a Bayesian must always condition on all known evidence, on pain of paradox. But then the clever arguer can make you believe anything he chooses, if there is a sufficient variety of signs to selectively report. That doesn't sound right.

Consider a simpler case, a biased coin, which may be biased to 2/3 heads 1/3 tails, or 1/3 heads 2/3 tails, both cases being equally likely a priori. Each H observed is 1 bit[3] of evidence for an H-biased coin; each T observed is 1 bit of evidence for a T-biased coin. I flip the coin ten times, and then I tell you, "The 4th flip, 6th flip, and 9th flip came up heads." What is your posterior probability that the coin is H-biased?

---

1. *http://lesswrong.com/lw/jt/what_evidence_filtered_evidence/*
2. Page 343, 'The Bottom Line'.
3. Page 22, 'How Much Evidence Does It Take?'.

And the answer is that it could be almost anything, depending on what chain of cause and effect lay behind my utterance of those words—my selection of which flips to report.

- I might be following the algorithm of reporting the result of the 4th, 6th, and 9th flips, regardless of the result of that and all other flips. If you know that I used this algorithm, the posterior odds are 8:1 in favor of an H-biased coin.
- I could be reporting on all flips, and only flips, that came up heads. In this case, you know that all 7 other flips came up tails, and the posterior odds are 1:16 against the coin being H-biased.
- I could have decided in advance to say the result of the 4th, 6th, and 9th flips only if the probability of the coin being H-biased exceeds 98%. And so on.

Or consider the Monty Hall problem:

> On a game show, you are given the choice of three doors leading to three rooms. You know that in one room is $100,000, and the other two are empty. The host asks you to pick a door, and you pick door #1. Then the host opens door #2, revealing an empty room. Do you want to switch to door #3, or stick with door #1?

The answer depends on the host's algorithm. If the host always opens a door and always picks a door leading to an empty room, then you should switch to door #3. If the host always opens door #2 regardless of what is behind it, #1 and #3 both have 50% probabilities of containing the money. If the host only opens a door, at all, if you initially pick the door with the money, then you should definitely stick with #1.

You shouldn't just condition on #2 being empty, but this fact plus the fact of the host *choosing* to open door #2. Many people are confused by the standard Monty Hall problem because they update only on #2 being empty, in which case #1 and #3

have equal probabilities of containing the money. This is why Bayesians are commanded to condition on all of their knowledge, on pain of paradox.

When someone says, "The 4th coinflip came up heads", we are not conditioning on the 4th coinflip having come up heads—we are not taking the subset of all possible worlds where the 4th coinflip came up heads—rather we are conditioning on the subset of all possible worlds where a speaker following some particular algorithm *said* "The 4th coinflip came up heads." The spoken sentence is not the fact itself; don't be led astray by the mere meanings of words.

Most legal processes work on the theory that every case has exactly two opposed sides[4] and that it is easier to find two biased humans than one unbiased one. Between the prosecution and the defense, *someone* has a motive to present any given piece of evidence, so the court will see all the evidence; that is the theory. If there are two clever arguers in the box dilemma, it is not quite as good as one curious inquirer, but it is almost as good. But that is with two boxes. Reality often has many-sided problems, and deep problems, and nonobvious answers, which are not readily found by Blues and Greens[5] screaming at each other[6].

Beware lest you abuse the notion of evidence-filtering as a Fully General Counterargument to exclude all evidence you don't like: "That argument was filtered, therefore I can ignore it." If you're ticked off by a contrary argument, then you are familiar with the case, and care enough to take sides. You probably already know your own side's strongest arguments. You have no reason to infer, from a contrary argument, the existence of new[7] favorable signs and portents which you have not

4. Page 154, 'The Scales of Justice, the Notebook of Rationality'.
5. Page 143, 'A Fable of Science and Politics'.
6. Page 148, 'Politics is the Mind-Killer'.
7. Page 340, 'One Argument Against An Army'.

yet seen[8]. So you are left with the uncomfortable facts themselves; a blue stamp on box B is still evidence.

But if you are hearing an argument for the first time, and you are only hearing one side of the argument, then indeed you should beware! In a way, no one can *really* trust the theory of natural selection until after they have listened to creationists for five minutes; and *then* they know it's solid.

8. Page 340, 'One Argument Against An Army'.

6. Rationalization[1]

Followup to: The Bottom Line[2], What Evidence Filtered Evidence?[3]

In "The Bottom Line", I presented the dilemma of two boxes only one of which contains a diamond, with various signs and portents as evidence. I dichotomized the curious inquirer and the clever arguer. The curious inquirer writes down all the signs and portents, and processes them, and finally writes down "*Therefore,* I estimate an 85% probability that box B contains the diamond." The clever arguer works for the highest bidder, and begins by writing, "*Therefore,* box B contains the diamond", and then selects favorable signs and portents to list on the lines above.

The first procedure is rationality. The second procedure is generally known as "rationalization".

"Rationalization." What a curious term. I would call it a *wrong word.* You cannot "rationalize" what is not already rational. It is as if "lying" were called "truthization".

On a purely computational level, there is a rather large difference between:

1. Starting from evidence, and then crunching probability flows, in order to output a probable conclusion. (Writing down all the signs and portents, and then flowing forward to a probability on the bottom line[4] which depends on those signs and portents.)
2. Starting from a conclusion, and then crunching probability flows, in order to output evidence apparently favoring that conclusion. (Writing down

1. *http://lesswrong.com/lw/ju/rationalization/*
2. Page 343, 'The Bottom Line'.
3. Page 347, 'What Evidence Filtered Evidence?'.
4. Page 343, 'The Bottom Line'.

> the bottom line, and then flowing backward to select[5] signs and portents for presentation[6] on the lines above.)

What fool devised such confusingly similar words, "rationality" and "rationalization", to describe such extraordinarily different mental processes? I would prefer terms that made the algorithmic difference obvious, like "rationality" versus "giant sucking cognitive black hole".

Not every change is an improvement, but every improvement is necessarily a change. You cannot obtain more truth for a fixed proposition by arguing it; you can make more people believe it, but you cannot make it more *true*. To improve our beliefs, we must necessarily change our beliefs. Rationality is the operation that we use to obtain more truth-value for our beliefs by changing them. Rationalization operates to fix beliefs in place; it would be better named "anti-rationality", both for its pragmatic results and for its reversed algorithm.

"Rationality" is the *forward* flow that gathers evidence, weighs it, and outputs a conclusion. The curious inquirer used a forward-flow algorithm: *first* gathering the evidence, writing down a list of all visible signs and portents, which they then processed *forward* to obtain a previously unknown probability for the box containing the diamond. During the entire time that the rationality-process was running forward, the curious inquirer did not yet know their destination, which was why they were *curious*. In the Way of Bayes, the prior probability equals the expected posterior probability[7]: If you know your destination, you are already there.

"Rationalization" is a *backward* flow from conclusion to selected evidence. First you write down the bottom line, which is known and fixed; the purpose of your processing is to find out which arguments you should write down on the lines above.

5. Page 347, 'What Evidence Filtered Evidence?'.
6. Page 347, 'What Evidence Filtered Evidence?'.
7. Page 68, 'Conservation of Expected Evidence'.

This, not the bottom line, is the variable unknown to the running process.

I fear that Traditional Rationality does not properly sensitize its users to the difference between forward flow and backward flow. In Traditional Rationality, there is nothing wrong with the scientist who arrives at a pet hypothesis and then sets out to find an experiment that proves it. A Traditional Rationalist would look at this approvingly, and say, "This pride is the engine that drives Science forward." Well, it *is* the engine that drives Science forward. It is easier to find a prosecutor and defender biased in opposite directions, than to find a single unbiased human.

But just because everyone does something, doesn't make it okay. It would be better yet if the scientist, arriving at a pet hypothesis, set out to *test* that hypothesis for the sake of *curiosity*—creating experiments that would drive their own beliefs in an unknown direction[8].

If you genuinely don't know where you are going, you will probably feel quite curious about it. Curiosity is the first virtue[9], without which your questioning will be purposeless and your skills without direction.

Feel the flow of the Force, and make sure it isn't flowing backwards.

8. Page 68, 'Conservation of Expected Evidence'.
9. *http://yudkowsky.net/virtues/*

7. A Rational Argument[1]

Followup to: The Bottom Line[2], Rationalization[3]

You are, by occupation, a campaign manager, and you've just been hired by Mortimer Q. Snodgrass, the Green candidate for Mayor of Hadleyburg. As a campaign manager reading a blog on rationality, one question lies foremost on your mind: "How can I construct an impeccable rational argument that Mortimer Q. Snodgrass is the best candidate for Mayor of Hadleyburg?"

Sorry. It can't be done.

"What?" you cry. "But what if I use only valid support to construct my structure of reason? What if every fact I cite is true to the best of my knowledge, and relevant evidence[4] under Bayes's Rule[5]?"

Sorry. It still can't be done. You defeated yourself the instant you specified your argument's conclusion in advance.

This year, the *Hadleyburg Trumpet* sent out a 16-item questionnaire to all mayoral candidates, with questions like "Can you paint with all the colors of the wind?" and "Did you inhale?" Alas, the *Trumpet's* offices are destroyed by a meteorite before publication. It's a pity, since your own candidate, Mortimer Q. Snodgrass, compares well to his opponents on 15 out of 16 questions. The only sticking point was Question 11, "Are you now, or have you ever been, a supervillain?"

So you are tempted to publish the questionnaire as part of your own campaign literature... with the 11th question omitted, of course.

Which crosses the line between *rationality* and *rationalization*. It is no longer possible for the voters to condition on the

1. *http://lesswrong.com/lw/jw/a_rational_argument/*
2. Page 343, 'The Bottom Line'.
3. Page 351, 'Rationalization'.
4. Page 18, 'What is Evidence?'.
5. *http://yudkowsky.net/rational/bayes*

facts alone; they must condition on the additional fact[6] of their presentation, and infer the existence of hidden evidence.

Indeed, you crossed the line at the point where you considered whether the questionnaire was favorable or unfavorable to your candidate, before deciding whether to publish it. "What!" you cry. "A campaign should publish facts unfavorable to their candidate?" But put yourself in the shoes of a voter, still trying to select a candidate—why would you censor useful information? You wouldn't, if you were genuinely curious. If you were flowing *forward* from the evidence to an unknown choice of candidate, rather than flowing *backward* from a fixed candidate to determine the arguments.

A "logical" argument is one that follows from its premises. Thus the following argument is *illogical:*

- All rectangles are quadrilaterals.
- All squares are quadrilaterals.
- *Therefore,* all squares are rectangles.

This syllogism is not rescued from illogic by the truth of its premises or even the truth of its conclusion. It is worth distinguishing logical deductions from illogical ones, and to refuse to excuse them even if their conclusions happen to be true. For one thing, the distinction may affect how we revise our beliefs in light of future evidence. For another, sloppiness is habit-forming.

Above all, the syllogism fails to state the real explanation. Maybe all squares are rectangles, but, if so, it's not *because* they are both quadrilaterals. You might call it a hypocritical syllogism—one with a disconnect between its stated reasons and real reasons.

If you really want to present an honest, rational argument *for your candidate,* in a political campaign, there is only one way to do it:

- *Before anyone hires you,* gather up all the evidence you can about the different candidates.

6. Page 347, 'What Evidence Filtered Evidence?'.

- Make a checklist which you, yourself, will use to decide which candidate seems best.
- Process the checklist.
- Go to the winning candidate.
- Offer to become their campaign manager.
- When they ask for campaign literature, print out your checklist.

Only in this way can you offer a *rational* chain of argument, one whose bottom line[7] was written flowing *forward* from the lines above it. Whatever *actually* decides your bottom line, is the only thing you can *honestly* write on the lines above.

7. Page 343, 'The Bottom Line'.

8. Avoiding Your Belief's Real Weak Points[1]

A few years back, my great-grandmother died, in her nineties, after a long, slow, and cruel disintegration. I never knew her as a person, but in my distant childhood, she cooked for her family; I remember her gefilte fish, and her face, and that she was kind to me. At her funeral, my grand-uncle, who had taken care of her for years, spoke: He said, choking back tears, that God had called back his mother piece by piece: her memory, and her speech, and then finally her smile; and that when God finally took her smile, he knew it wouldn't be long before she died, because it meant that she was almost entirely gone.

I heard this and was puzzled, because it was an unthinkably horrible thing to happen to *anyone,* and therefore I would not have expected my grand-uncle to attribute it to God. Usually, a Jew would somehow just-not-think-about the logical implication that God had permitted a tragedy. According to Jewish theology, God continually sustains the universe and chooses every event in it; but ordinarily, drawing logical implications from this belief is reserved for happier occasions. By saying "God did it!" only when you've been blessed with a baby girl, and just-not-thinking "God did it!" for miscarriages and stillbirths and crib deaths, you can build up quite a lopsided[2] picture of your God's benevolent personality.

Hence I was surprised to hear my grand-uncle attributing the slow disintegration of his mother to a deliberate, strategically planned act of God. It violated the rules of religious self-deception as I understood them.

If I had noticed my own confusion[3], I could have made a successful surprising prediction. Not long afterward, my

1. *http://lesswrong.com/lw/jy/avoiding_your_beliefs_real_weak_points/*
2. Page 351, 'Rationalization'.
3. Page 62, 'Your Strength as a Rationalist'.

grand-uncle left the Jewish religion. (The only member of my extended family besides myself to do so, as far as I know.)

Modern Orthodox Judaism[4] is like no other religion I have ever heard of, and I don't know how to describe it to anyone who hasn't been forced to study Mishna and Gemara. There is a tradition of questioning, but the *kind* of questioning... It would not be at all surprising to hear a rabbi, in his weekly sermon, point out the conflict between the seven days of creation and the 13.7 billion years since the Big Bang—because he thought he had a really clever explanation for it, involving three other Biblical references, a Midrash, and a half-understood article in *Scientific American*. In Orthodox Judaism you're allowed to notice inconsistencies and contradictions, but only for purposes of explaining them away, and whoever comes up with the most complicated explanation gets a prize.

There is a tradition of inquiry. But you only attack targets for purposes of defending them. You only attack targets you know you can defend.

In Modern Orthodox Judaism I have not heard much emphasis of the virtues of blind faith. You're allowed to doubt. You're just not allowed to *successfully* doubt[5].

I expect that the vast majority of educated Orthodox Jews have questioned their faith at some point in their lives. But the questioning probably went something like this: "According to the skeptics, the Torah says that the universe was created in seven days, which is not scientifically accurate. But would the original tribespeople of Israel, gathered at Mount Sinai, have been able to understand the scientific truth, even if it had been presented to them? Did they even have a word for 'billion'? It's easier to see the seven-days story as a metaphor—first God created light, which represents the Big Bang..."

Is this the weakest point at which to attack one's own Judaism? Read a bit further on in the Torah, and you can find

4. *http://en.wikipedia.org/wiki/Modern_Orthodox_Judaism*
5. Page 474, 'The Proper Use of Doubt'.

God killing the first-born male children of Egypt to convince an unelected Pharaoh to release slaves who logically could have been teleported out of the country. An Orthodox Jew is most certainly familiar with this episode, because they are supposed to read through the entire Torah in synagogue once per year, and this event has an associated major holiday. The name "Passover" ("Pesach") comes from God *passing over* the Jewish households while killing every male firstborn in Egypt.

Modern Orthodox Jews are, by and large, kind and civilized people; far more civilized than the several editors of the Old Testament. Even the old rabbis were more civilized. There's a ritual in the Seder where you take ten drops of wine from your cup, one drop for each of the Ten Plagues, to emphasize the suffering of the Egyptians. (Of course, you're supposed to be sympathetic to the suffering of the Egyptians, but not *so* sympathetic that you stand up and say, "This is not right! It is *wrong* to do such a thing!") It shows an interesting contrast—the rabbis were sufficiently kinder than the compilers of the Old Testament that they saw the harshness of the Plagues. But Science was weaker in these days, and so rabbis could ponder the more unpleasant aspects of Scripture without fearing that it would break their faith entirely.

You don't even *ask* whether the incident reflects poorly on God, so there's no need to quickly blurt out "The ways of God are mysterious!" or "We're not wise enough to question God's decisions!" or "Murdering babies is okay when God does it!" That part of the question is just-not-thought-about.

The reason that educated religious people stay religious, I suspect, is that when they doubt, they are subconsciously very careful to attack their own beliefs only at the strongest points—places where they know they can defend. Moreover, places where rehearsing[6] the standard defense will feel strengthening.

6. Page 340, 'One Argument Against An Army'.

It probably feels really good, for example, to rehearse one's prescripted defense for "Doesn't Science say that the universe is just meaningless atoms bopping around?", because it confirms the meaning of the universe and how it flows from God, etc.. Much more comfortable to think about than an illiterate Egyptian mother wailing over the crib of her slaughtered son. Anyone who *spontaneously* thinks about the latter, when questioning their faith in Judaism, is *really* questioning it, and is probably not going to stay Jewish much longer.

My point here is not just to beat up on Orthodox Judaism. I'm sure that there's some reply or other for the Slaying of the Firstborn, and probably a dozen of them. My point is that, when it comes to spontaneous self-questioning, one is much more likely to spontaneously self-attack strong points with comforting replies to rehearse, then to spontaneously self-attack the weakest, most vulnerable points. Similarly, one is likely to stop at the first reply and be comforted, rather than further criticizing the reply. A better title than "Avoiding Your Belief's Real Weak Points" would be "Not Spontaneously Thinking About Your Belief's Most Painful Weaknesses".

More than anything, the grip of religion is sustained by people just-not-thinking-about the real weak points of their religion. I don't think this is a matter of training, but a matter of instinct. People don't think about the real weak points of their beliefs for the same reason they don't touch an oven's red-hot burners; it's *painful*.

To do better[7]: When you're doubting one of your most cherished beliefs, close your eyes, empty your mind, grit your teeth, and deliberately think about whatever hurts the most. Don't rehearse standard objections whose standard counters would make you feel better. Ask yourself what *smart* people who disagree would say to your first reply, and your second reply. Whenever you catch yourself flinching away from an objection you fleetingly thought of, drag it out into the forefront of your

7. *http://lesswrong.com/lw/h8/tsuyoku_naritai_i_want_to_become_stronger/*

mind. Punch yourself in the solar plexus. Stick a knife in your heart, and wiggle to widen the hole. In the face of the pain, rehearse only this:

What is true is already so.
Owning up to it doesn't make it worse.
Not being open about it doesn't make it go away.
And because it's true, it is what is there to be interacted with.
Anything untrue isn't there to be lived.
People can stand what is true,
for they are already enduring it.
—*Eugene Gendlin*

9. Motivated Stopping and Motivated Continuation[1]

Followup to: The Third Alternative[2], The Meditation on Curiosity[3]

While I disagree with some views of the Fast and Frugal[4] crowd—IMO they make a few *too* many lemons into lemonade—it also seems to me that they tend to develop the most *psychologically realistic* models of any school of decision theory. Most experiments present the subjects with options, and the subject chooses an option, and that's the experimental result. The frugalists realized that in real life, you have to *generate* your options, and they studied how subjects did *that*.

Likewise, although many experiments present evidence on a silver platter, in real life you have to gather evidence, which may be costly, and at some point decide that you have enough evidence to stop and choose. When you're buying a house, you don't get exactly 10 houses to choose from, and you aren't led on a guided tour of all of them before you're allowed to decide anything. You look at one house, and another, and compare them to each other; you adjust your aspirations—reconsider how much you really need to be close to your workplace and how much you're really willing to pay; you decide which house to look at next; and at some point you decide that you've seen enough houses, and choose.

Gilovich's distinction between *motivated skepticism* and *motivated credulity* highlights how conclusions a person does not want to believe are held to a higher standard than conclusions a person wants to believe. A motivated skeptic asks if the evidence *compels* them to accept the conclusion; a motivated

1. *http://lesswrong.com/lw/km/motivated_stopping_and_motivated_continuation/*
2. Page 426, 'The Third Alternative'.
3. Page 478, 'The Meditation on Curiosity'.
4. *http://fastandfrugal.com/*

credulist asks if the evidence *allows* them to accept the conclusion.

I suggest that an analogous bias in psychologically realistic search is *motivated stopping* and *motivated continuation:* when we have a *hidden* motive for choosing the "best" current option, we have a hidden motive to stop, and choose, and reject consideration of any more options. When we have a hidden motive to reject the current best option, we have a hidden motive to suspend judgment pending additional evidence, to generate more options—to find something, anything, to do *instead* of coming to a conclusion.

A major historical scandal in statistics was R. A. Fisher, an eminent founder of the field, insisting that no *causal* link had been established between smoking and lung cancer. "Correlation is not causation", he testified to Congress. Perhaps smokers had a gene which both predisposed them to smoke and predisposed them to lung cancer.

Or maybe Fisher being employed as a consultant for tobacco firms gave him a hidden motive to decide that the evidence already gathered was insufficient to come to a conclusion, and it was better to keep looking. Fisher was also a smoker himself, and died of colon cancer in 1962.

(Ad hominem note: Fisher was a frequentist. Bayesians[5] are more reasonable about inferring probable causality.)

Like many other forms of motivated skepticism, motivated continuation can try to disguise itself as virtuous rationality. Who can argue against gathering more evidence [6]? I can. Evidence is often costly, and worse, slow, and there is certainly nothing virtuous about refusing to integrate the evidence you already have. You can always change your mind later.[7] (Apparent contradiction resolved as follows: Spending *one hour* discussing the problem with your mind carefully cleared of all

5. *http://bayes.cs.ucla.edu/BOOK-2K/*
6. *http://www.overcomingbias.com/2007/01/conspicuous_con.html*
7. Page 320, 'Hold Off On Proposing Solutions'.

conclusions, is different from waiting ten years on another $20 million study.)

As for motivated stopping, it appears in every place a third alternative[8] is feared, and wherever you have an argument whose obvious counterargument[9] you would rather not see, and in other places as well. It appears when you pursue a course of action that makes you feel good just for acting[10], and so you'd rather not investigate how well your plan *really* worked, for fear of destroying the warm glow of moral satisfaction[11] you paid good money to purchase. It appears wherever your beliefs and anticipations get out of sync[12], so you have a reason to fear any new evidence gathered.

The moral is that the decision to terminate a search procedure (temporarily or permanently) is, like the search procedure itself, subject to bias and hidden motives. You should suspect motivated stopping when you close off search, after coming to a comfortable conclusion, and yet there's a lot of fast cheap evidence you haven't gathered yet—Web sites you could visit, counter-counter arguments you could consider, or you haven't closed your eyes for five minutes by the clock trying to think of a better option. You should suspect motivated continuation when some evidence is leaning in a way you don't like, but you decide that more evidence is needed—*expensive* evidence that you know you can't gather anytime soon, as opposed to something you're going to look up on Google in 30 minutes—before you'll have to do anything uncomfortable.

8. Page 426, 'The Third Alternative'.
9. Page 478, 'The Meditation on Curiosity'.
10. *http://lesswrong.com/lw/kb/cant_say_no_spending/*
11. *http://lesswrong.com/lw/hw/scope_insensitivity/*
12. Page 43, 'Belief in Belief'.

10. A Case Study of Motivated Continuation[1]

I am not wholly unsympathetic to the many commenters in Torture vs. Dust Specks[2] who argued that it is preferable to inflict dust specks upon the eyes of 3^^^3 (amazingly huge but finite number of) people, rather than torture one person for 50 years. If you think that a dust speck is simply of no account unless it has other side effects - if you literally do not prefer zero dust specks to one dust speck - then your position is consistent. (Though I suspect that many speckers would have expressed a preference if they hadn't known about the dilemma's sting.)

So I'm on board with the commenters who chose TORTURE, and I can understand the commenters who chose SPECKS.

But some of you said the question was meaningless; or that all morality was arbitrary and subjective; or that you needed more information before you could decide; or you talked about some other confusing aspect of the problem; and then you *didn't* go on to state a preference.

Sorry. I can't back you on that one.

If you actually answer the dilemma, then no matter which option you choose, you're giving something up. If you say SPECKS, you're giving up your claim on a certain kind of utilitarianism; you may worry that you're not being rational enough, or that others will accuse you of failing to comprehend large numbers. If you say TORTURE, you're accepting an outcome that has torture in it.

I falsifiably predict that of the commenters who dodged, most of them saw some specific answer - either TORTURE or SPECKS - that they flinched away from giving. Maybe for just a fraction of a second before the question-confusing operation

1. *http://lesswrong.com/lw/ko/a_case_study_of_motivated_continuation/*
2. *http://lesswrong.com/lw/kn/torture_vs_dust_specks/*

took over, but I predict the flinch was there. (To be specific: I'm not predicting that you knew, and selected, and have in mind right now, some particular answer you're deliberately not giving. I'm predicting that your thinking trended toward a particular uncomfortable answer, for at least one fraction of a second before you started finding reasons to question the dilemma itself.)

In "bioethics[3]" debates, you very often see experts on[4] bioethics discussing what they see as the pros and cons of, say, stem-cell research; and then, at the conclusion of their talk, they gravely declare that more debate is urgently needed, with participation[5] from all stakeholders. If you actually come to a conclusion, if you actually argue for banning stem cells, then people with relatives dying of Parkinson's will scream at you. If you come to a conclusion and actually endorse stem cells, religious fundamentalists will scream at you. But who can argue with a call to debate[6]?

Uncomfortable with the way the evidence is trending on Darwinism versus creationism? Consider the issue soberly, and decide that you need more evidence; you want archaeologists to dig up another billion fossils before you come to a conclusion. That way you neither say something sacrilegious, nor relinquish your self-image as a rationalist. Keep on doing this with all issues that look like they might be trending in an uncomfortable direction, and you can maintain a whole religion in your mind.

Real life is often confusing, and we have to choose anyway, because refusing to choose is also a choice. The null plan is still a plan. We always do *something,* even if it's nothing. As Russell and Norvig put it, "Refusing to choose is like refusing to allow time to pass."

---

3. *http://intelligence.org/blog/2007/10/21/should-ethicists-be-inside-or-outside-a-profession/*
4. *http://www.overcomingbias.com/2007/04/expert_at_versu.html*
5. *http://lesswrong.com/lw/ja/we_dont_really_want_your_participation/*
6. Page 128, 'Applause Lights'.

Ducking uncomfortable choices is a dangerous habit of mind. There are certain times when it's wise to suspend judgment[7] (for an hour, not a year). When you're facing a dilemma all of whose answers seem uncomfortable, is *not* one of those times! Pick *one* of the uncomfortable answers as the best of an unsatisfactory lot. If there's missing information, fill in the blanks with plausible assumptions or probability distributions. Whatever it takes to overcome the basic flinch away from discomfort. *Then* you can search for an escape route[8].

Until you pick one interim best guess, the discomfort will consume your attention, distract you from the search, tempt you to confuse the issue whenever your analysis seems to trend in a particular direction.

In real life, when people flinch away from uncomfortable choices, they often hurt others as well as themselves. Refusing to choose is often one of the worst choices you can make. Motivated continuation[9] is not a habit of thought anyone can afford, egoist or altruist. The cost of comfort is too high. It's important to acquire that habit of gritting your teeth and choosing - just as important as looking for escape routes *afterward*.

7. Page 320, 'Hold Off On Proposing Solutions'.
8. Page 426, 'The Third Alternative'.
9. Page 362, 'Motivated Stopping and Motivated Continuation'.

11. Fake Justification[1]

Many Christians who've stopped really believing[2] now insist that they revere the Bible as a source of ethical advice. The standard atheist reply is given by Sam Harris[3]: "You and I both know that it would take us five minutes to produce a book that offers a more coherent and compassionate morality than the Bible does." Similarly, one may try to insist that the Bible is valuable as a literary work. Then why not revere *Lord of the Rings,* a vastly superior literary work? And despite the standard criticisms of Tolkien's morality, *Lord of the Rings* is at least superior to the Bible as a source of ethics. So why don't people wear little rings around their neck, instead of crosses? Even *Harry Potter* is superior to the Bible, both as a work of literary art and as moral philosophy. If I really wanted to be cruel, I would compare the Bible to Jacqueline Carey's *Kushiel* series.

"How can you justify buying a $1 million gem-studded laptop[4]," you ask your friend, "when so many people have no laptops at all?" And your friend says, "But think of the employment that this will provide—to the laptop maker, the laptop maker's advertising agency—and then they'll buy meals and haircuts—it will stimulate the economy and eventually many people will get their own laptops." But it would be even *more* efficient to buy 5,000 OLPC laptops, thus providing employment to the OLPC manufacturers *and* giving out laptops directly.

I've touched before on the failure to look for third alternatives[5]. But this is not really motivated stopping[6]. Calling it

1. *http://lesswrong.com/lw/kq/fake_justification/*
2. Page 43, 'Belief in Belief'.
3. *http://homepage.mac.com/pmcarlton/Harris_Sullivan_CompleteDebate.pdf*
4. *http://hardware.slashdot.org/article.pl?sid=07/03/26/197253*
5. Page 426, 'The Third Alternative'.
6. Page 362, 'Motivated Stopping and Motivated Continuation'.

"motivated stopping" would imply that there was a search carried out in the first place.

In The Bottom Line[7], I observed that only the real determinants of our beliefs can ever influence our real-world accuracy, only the real determinants of our actions can influence our effectiveness in achieving our goals. Someone who buys a million-dollar laptop was really thinking, "Ooh, shiny" and that was the one true causal history of their decision to buy a laptop. No amount of "justification" can change this, unless the justification is a genuine, newly running search process that can change the conclusion. *Really* change the conclusion. Most criticism carried out from a sense of duty[8] is more of a token inspection than anything else. Free elections in a one-party country.

To genuinely justify the Bible as a lauding-object by reference to its literary quality, you would have to somehow perform a neutral reading through candidate books until you found the book of highest literary quality. Renown is one reasonable criteria for generating candidates, so I suppose you could legitimately end up reading Shakespeare, the Bible, and *Godel, Escher, Bach*. (Otherwise it would be quite a coincidence to find the Bible as a candidate, among a million other books.) The real difficulty is in that "neutral reading" part. Easy enough if you're not a Christian, but if you are...

But of course nothing like this happened. No search ever occurred. Writing the justification of "literary quality" above the bottom line[9] of "I <heart> the Bible" is a historical misrepresentation of how the bottom line[10] really got there, like selling cat milk as cow milk. That is just not where the bottom line[11]

7. Page 343, 'The Bottom Line'.
8. Page 478, 'The Meditation on Curiosity'.
9. Page 343, 'The Bottom Line'.
10. Page 343, 'The Bottom Line'.
11. Page 343, 'The Bottom Line'.

really came from. That is just not what originally happened to produce that conclusion.

If you genuinely subject your conclusion to a criticism that can potentially de-conclude it—if the criticism *genuinely* has that power—then that does modify "the real algorithm behind" your conclusion. It changes the entanglement of your conclusion over possible worlds. But people overestimate, by far, how likely they *really* are to change their minds[12].

With all those open minds out there, you'd think there'd be more belief-updating.

Let me guess: Yes, you admit that you originally decided you wanted to buy a million-dollar laptop by thinking, "Ooh, shiny". Yes, you concede that this isn't a decision process consonant with your stated goals. But since then, you've decided that you really ought to spend your money in such fashion as to provide laptops to as many laptopless wretches as possible. And yet you just *couldn't* find any more efficient way to do this than buying a million-dollar diamond-studded laptop—because, hey, you're giving money to a laptop store and stimulating the economy! Can't beat that!

My friend, I am damned suspicious of this amazing coincidence. I am damned suspicious that the best answer under this lovely, rational, altruistic criterion X, is also the idea that just happened to originally pop out of the unrelated indefensible process Y. If you don't think that rolling dice would have been likely to produce the correct answer, then how likely is it to pop out of any other irrational cognition?

It's improbable that you used mistaken reasoning, yet made no mistakes.

12. Page 318, 'We Change Our Minds Less Often Than We Think'.

12. Fake Optimization Criteria[1]

Followup to: Fake Justification[2], The Tragedy of Group Selectionism[3]

I've[4] previously[5] dwelt[6] in[7] considerable[8] length[9] upon[10] forms[11] of[12] rationalization[13] whereby[14] our[15] beliefs[16] appear [17]to[18] match[19] the[20] evidence[21] much[22] more[23] strongly[24] than[25] they[26] actually[27] do[28]. And I'm not overemphasizing the point, either. If we could beat this fundamental metabias and see what every hypothesis *really* predicted, we would be able to recover from almost any other error of fact.

The mirror challenge for decision theory is seeing which option a choice criterion *really* endorses. If your stated moral

1. *http://lesswrong.com/lw/kz/fake_optimization_criteria/*
2. Page 368, 'Fake Justification'.
3. *http://lesswrong.com/lw/kw/the_tragedy_of_group_selectionism/*
4. Page 71, 'Hindsight bias'.
5. Page 74, 'Hindsight Devalues Science'.
6. Page 62, 'Your Strength as a Rationalist'.
7. Page 55, 'Focus Your Uncertainty'.
8. Page 65, 'Absence of Evidence Is Evidence of Absence'.
9. Page 68, 'Conservation of Expected Evidence'.
10. Page 39, 'Making Beliefs Pay Rent (in Anticipated Experiences)'.
11. Page 43, 'Belief in Belief'.
12. Page 347, 'What Evidence Filtered Evidence?'.
13. Page 351, 'Rationalization'.
14. Page 50, 'Professing and Cheering'.
15. Page 53, 'Belief as Attire'.
16. *http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/*
17. Page 77, 'Fake Explanations'.
18. Page 80, 'Guessing the Teacher's Password'.
19. Page 87, 'Fake Causality'.
20. Page 92, 'Semantic Stopsigns'.
21. Page 18, 'What is Evidence?'.
22. Page 96, 'Mysterious Answers to Mysterious Questions'.
23. Page 100, 'The Futility of Emergence'.
24. Page 104, 'Say Not "Complexity"'.
25. Page 108, 'Positive Bias: Look Into the Dark'.
26. Page 333, 'Knowing About Biases Can Hurt People'.
27. Page 343, 'The Bottom Line'.
28. Page 347, 'What Evidence Filtered Evidence?'.

principles[29] call for you to provide laptops to everyone, does that *really* endorse buying a $1 million gem-studded laptop for yourself, or spending the same money on shipping 5000 OLPCs?

We seem to have evolved a knack for arguing that practically any goal implies practically any action. A phlogiston theorist explaining why magnesium gains weight when burned has nothing on an Inquisitor explaining why God's infinite love for all His children requires burning some of them at the stake.

There's no mystery about this. Politics[30] was a feature of the ancestral environment. We are descended from those who argued most persuasively that the good of the tribe meant executing their hated rival Uglak. (We sure ain't descended from Uglak.)

And yet... is it possible to *prove* that if Robert Mugabe cared *only* for the good of Zimbabwe, he would resign from its presidency? You can *argue* that the policy follows from the goal, but haven't we just seen that humans can match up any goal to any policy? How do you know that you're right and Mugabe is wrong? (There are a number of reasons this is a good guess, but bear with me here.)

Human motives are manifold and obscure, our decision processes as vastly complicated as our brains. And the world itself is vastly complicated, on every choice of real-world policy. Can we even *prove* that human beings are rationalizing—that we're systematically distorting the link from principles to policy—when we lack a single firm place on which to stand? When there's no way to find out *exactly* what even a single optimization criterion implies? (Actually, you can just observe that people *disagree* about office politics in ways that strangely correlate to their own interests, while simultaneously denying that any such interests are at work. But again, bear with me here.)

29. Page 368, 'Fake Justification'.
30. Page 148, 'Politics is the Mind-Killer'.

Where is the standardized, open-source, generally intelligent, consequentialist optimization process into which we can feed a complete morality as an XML file, to find out what that morality *really* recommends when applied to our world? Is there even a single real-world case where we can know *exactly* what a choice criterion recommends? Where is the *pure* moral reasoner—of known utility function, purged of all other stray desires that might distort its optimization—whose trustworthy output we can contrast to human rationalizations of the same utility function?

Why, it's our old friend the alien god[31], of course! Natural selection is guaranteed free of all mercy, all love, all compassion, all aesthetic sensibilities, all political factionalism, all ideological allegiances, all academic ambitions, all libertarianism, all socialism, all Blue and all Green[32]. Natural selection doesn't *maximize* its criterion of inclusive genetic fitness—it's not that smart[33]. But when you look at the output of natural selection, you are guaranteed to be looking at an output that was optimized *only* for inclusive genetic fitness, and not the interests of the US agricultural industry.

In the case histories of evolutionary science—in, for example, The Tragedy of Group Selectionism[34]—we can directly compare human rationalizations to the result of *pure* optimization for a known criterion. What did Wynne-Edwards think would be the result of group selection for small subpopulation sizes? Voluntary individual restraint in breeding, and enough food for everyone. What was the actual laboratory result? Cannibalism.

Now you might ask: Are these case histories of evolutionary science really relevant to human morality, which doesn't give two figs for inclusive genetic fitness when it gets in the way

31. *http://lesswrong.com/lw/kr/an_alien_god/*
32. Page 143, 'A Fable of Science and Politics'.
33. *http://lesswrong.com/lw/kt/evolutions_are_stupid_but_work_anyway/*
34. *http://lesswrong.com/lw/kw/the_tragedy_of_group_selectionism/*

of love, compassion, aesthetics, healing, freedom, fairness, et cetera? Human societies didn't even have a concept of "inclusive genetic fitness" until the 20th century.

But I ask in return: If we can't see clearly the result of a single monotone optimization criterion—if we can't even train ourselves to hear a single pure note—then how will we listen to an orchestra? How will we see that "Always be selfish" or "Always obey the government" are poor guiding principles for human beings to adopt—if we think that even *optimizing genes for inclusive fitness* will yield organisms which sacrifice reproductive opportunities in the name of social resource conservation?

To train ourselves to see clearly, we need simple practice cases.

(end of *The Simple Math of Evolution*[35])

35. *http://wiki.lesswrong.com/wiki/Evolution#Blog_posts_.28sequence.29*

13. Is That Your True Rejection?[1]

It happens every now and then, that the one encounters some of my transhumanist-side beliefs—as opposed to my ideas having to do with human rationality—strange, exotic-sounding ideas like superintelligence and Friendly AI. And the one rejects them.

If the one is called upon to explain the rejection, not uncommonly the one says,

"Why should I believe anything Yudkowsky says? He doesn't have a PhD!"

And occasionally someone else, hearing, says, "Oh, you should get a PhD, so that people will listen to you." Or this advice may even be offered by the same one who disbelieved, saying, "Come back when you have a PhD."

Now there are good and bad reasons to get a PhD, but this is one of the bad ones.

There's many reasons why someone *actually* has an adverse reaction to transhumanist theses. Most are matters of pattern recognition, rather than verbal thought: the thesis matches[2] against "strange weird idea" or "science fiction" or "end-of-the-world cult" or "overenthusiastic youth".

So immediately, at the speed of perception, the idea is rejected. If, afterward, someone says "Why not?", this lanches a search for justification. But this search will not necessarily hit on the true reason—by "true reason" I mean not the *best* reason that could be offered, but rather, whichever causes were decisive as a matter of historical fact[3], at the *very first* moment the rejection occurred[4].

---

1. *http://lesswrong.com/lw/wj/is_that_your_true_rejection/*
2. Page 84, 'Science as Attire'.
3. Page 343, 'The Bottom Line'.
4. Page 318, 'We Change Our Minds Less Often Than We Think'.

Instead, the search for justification hits on the justifying-sounding fact, "This speaker does not have a PhD."

But I also don't have a PhD when I talk about human rationality, so why is the same objection not raised there[5]?

And more to the point, if I *had* a PhD, people would not treat this as a decisive factor indicating that they ought to believe everything I say. Rather, the same initial rejection would occur, for the same reasons; and the search for justification, afterward, would terminate at a different stopping point.

They would say, "Why should I believe *you?* You're just some guy with a PhD! There are lots of those. Come back when you're well-known in your field and tenured at a major university."

But do people *actually* believe arbitrary professors at Harvard who say weird things? Of course not. (But if I were a professor at Harvard, it would in fact be easier to get *media attention*. Reporters initially disinclined to believe me—who would probably be equally disinclined to believe a random PhD-bearer—would still report on me, because it would be news that a Harvard professor believes such a weird thing.)

If you are saying things that sound *wrong* to a novice, as opposed to just rattling off magical-sounding technobabble about leptical quark braids in N+2 dimensions; and the hearer is a stranger, unfamiliar with you personally *and* with the subject matter of your field; then I suspect that the point at which the average person will *actually* start to grant credence overriding their initial impression, purely *because* of academic credentials, is somewhere around the Nobel Laureate level. If that. Roughly, you need whatever level of academic credential qualifies as "beyond the mundane".

This is more or less what happened to Eric Drexler, as far as I can tell. He presented his vision of nanotechnology, and people said, "Where are the technical details?" or "Come back when you have a PhD!" And Eric Drexler spent six years writing

5. Page 276, 'Cultish Countercultishness'.

up technical details and got his PhD under Marvin Minsky for doing it. And *Nanosystems* is a great book. But did the same people who said, "Come back when you have a PhD", actually change their minds at all about molecular nanotechnology? Not so far as I ever heard.

It has similarly been a general rule with the Singularity Institute that, whatever it is we're supposed to do to be more credible, when we actually do it, nothing much changes. "Do you do any sort of code development? I'm not interested in supporting an organization that doesn't develop code"—> OpenCog—> nothing changes. "Eliezer Yudkowsky lacks academic credentials"—> Professor Ben Goertzel installed as Director of Research—> nothing changes. The one thing that actually *has* seemed to raise credibility, is famous people associating with the organization, like Peter Thiel funding us, or Ray Kurzweil on the Board.

This might be an important thing for young businesses and new-minted consultants to keep in mind—that what your failed prospects *tell* you is the reason for rejection, may not make the *real* difference; and you should ponder that carefully before spending huge efforts. If the venture capitalist says "If only your sales were growing a little faster!", if the potential customer says "It seems good, but you don't have feature X", that may not be the *true* rejection. Fixing it may, or may not, change anything.

And it would also be something to keep in mind during disagreements. Robin and I share a belief that two rationalists should not agree to disagree[6]: they should not have common knowledge of epistemic disagreement unless something is very wrong.

I suspect that, in general, if two rationalists set out to resolve a disagreement that persisted past the first exchange, they should expect to find that the true sources of the disagreement are either hard to communicate, or hard to expose. E.g:

6. *http://www.overcomingbias.com/2006/12/agreeing_to_agr.html*

- Uncommon, but well-supported, scientific knowledge or math;
- Long inferential distances[7];
- Hard-to-verbalize intuitions, perhaps stemming from specific visualizations;
- Zeitgeists inherited from a profession (that may have good reason for it);
- Patterns perceptually recognized from experience;
- Sheer habits of thought;
- Emotional commitments to believing in a particular outcome;
- Fear of a past mistake being disproven;
- Deep self-deception for the sake of pride or other personal benefits.

If the matter were one in which *all* the true rejections could be *easily* laid on the table, the disagreement would probably be so straightforward to resolve that it would never have lasted past the first meeting.

"Is this my true rejection?" is something that both disagreers should surely be asking *themselves,* to make things easier on the Other Fellow. However, attempts to directly, publicly psychoanalyze the Other may cause the conversation to degenerate *very* fast, in my observation.

Still—"Is that your true rejection?" should be fair game for Disagreers to humbly ask, if there's any productive way to pursue that sub-issue. Maybe the rule could be that you can openly ask, "Is that simple straightforward-sounding reason your *true* rejection, or does it come from intuition-X or professional-zeitgeist-Y?" While the more embarrassing possibilities lower on the table are left to the Other's conscience, as their own responsibility to handle.

Post scriptum:

This post is not *really* about PhDs in general, or their credibility value in particular. But I've always figured that to the

7. *http://lesswrong.com/lw/kg/expecting_short_inferential_distances/*

extent this was a strategically important consideration, it would make more sense to recruit an academic of existing high status, than spend a huge amount of time trying to achieve low or moderate academic status.

However, if any professor out there wants to let me come in and *just* do a PhD in analytic philosophy—*just* write the thesis and defend it—then I have, for my own use, worked out a general and mathematically elegant theory of Newcomblike decision problems[8]. I think it would make a fine PhD thesis, and it is ready to be written—if anyone has the power to let me do things the old-fashioned way.

8. *http://lesswrong.com/lw/nc/newcombs_problem_and_regret_of_rationality/*

14. Entangled Truths, Contagious Lies[1]

> "One of your very early philosophers came to the conclusion that a fully competent mind, from a study of one fact or artifact belonging to any given universe, could construct or visualize that universe, from the instant of its creation to its ultimate end..."
>
> —*First Lensman*

> "If any one of you will concentrate upon one single fact, or small object, such as a pebble or the seed of a plant or other creature, for as short a period of time as one hundred of your years, you will begin to perceive its truth."
>
> —*Gray Lensman*

I am reasonably sure that a single pebble, taken from a beach of our own Earth, does not specify the continents and countries, politics and people of this Earth. Other planets in space and time, other Everett branches[2], would generate the same pebble. On the other hand, the identity of a single pebble would seem to include our laws of physics. In that sense the entirety of our Universe—*all* the Everett branches—would be implied by the pebble. (If, as seems likely, there are no truly free variables.)

So a single pebble probably does not imply our whole Earth. But a single pebble implies a very great deal. From the study of that single pebble you could see the laws of physics and all they imply. Thinking about those laws of physics, you can see that planets will form, and you can guess that the pebble came from such a planet. The internal crystals and molecular formations of the pebble formed under gravity, which tells you something

1. *http://lesswrong.com/lw/uw/entangled_truths_contagious_lies/*
2. *http://lesswrong.com/lw/r8/and_the_winner_is_manyworlds/*

about the planet's mass; the mix of elements in the pebble tells you something about the planet's formation.

I am not a geologist, so I don't know to which mysteries geologists are privy. But I find it very easy to imagine showing a geologist a pebble, and saying, "This pebble came from a beach at Half Moon Bay", and the geologist immediately says, "I'm confused[3]" or even "You liar". Maybe it's the wrong kind of rock, or the pebble isn't worn enough to be from a beach—I don't know pebbles well enough to guess the linkages and signatures by which I might be caught, which is the point.

"Only God can tell a truly plausible lie." I wonder if there was ever a religion that developed this as a proverb? I would (falsifiably) guess not: it's a rationalist sentiment, even if you cast it in theological metaphor. Saying "everything is interconnected to everything else, because God made the whole world and sustains it" may generate some nice warm n' fuzzy feelings during the sermon, but it doesn't get you very far when it comes to assigning pebbles to beaches.

A penny on Earth exerts a gravitational acceleration on the Moon of around $4.5 * 10^{-31}$ m/s^2, so in one sense it's not too far wrong to say that every event is entangled with its whole past light cone. And since inferences can propagate backward and forward through causal networks, *epistemic* entanglements can easily cross the borders of light cones. But I wouldn't want to be the forensic astronomer[4] who had to look at the Moon and figure out whether the penny landed heads or tails—the influence is far less than quantum uncertainty and thermal noise.

If you said "Everything is entangled with something else" or "Everything is inferentially entangled and some entanglements are much stronger than others", you might be really wise instead of just Deeply Wise[5].

3. Page 62, 'Your Strength as a Rationalist'.
4. *http://very.net/~nikolai/tb/coroner.html*
5. Page 314, 'How to Seem (and Be) Deep'.

Physically, each event is in some sense the sum of its whole past light cone, without borders or boundaries. But the list of *noticeable* entanglements is much shorter, and it gives you something like a network. This high-level regularity[6] is what I refer to when I talk about the Great Web of Causality.

I use these Capitalized Letters somewhat tongue-in-cheek, perhaps; but if anything at all is worth Capitalized Letters, surely the Great Web of Causality makes the list.

"Oh what a tangled web we weave, when first we practise to deceive," said Sir Walter Scott. Not *all* lies spin out of control—we don't live in so righteous a universe[7]. But it does occasionally happen, that someone lies about a fact, and then has to lie about an entangled fact, and then another fact entangled with that one:

> "Where were you?"
> "Oh, I was on a business trip."
> "What was the business trip about?"
> "I can't tell you that; it's proprietary negotiations with a major client."
> "Oh—they're letting you in on those? Good news! I should call your boss to thank him for adding you."
> "Sorry—he's not in the office right now..."

Human beings, who are not gods, often fail to *imagine* all the facts they would need to distort to tell a truly plausible lie. "God made me pregnant[8]" sounded a tad more likely in the old days before our models of the world contained (quotations of) Y chromosomes. Many similar lies, today, may blow up when genetic testing becomes more common. Rapists have been convicted, and false accusers exposed, years later, based on evidence they didn't realize they could leave. A student of

6. *http://lesswrong.com/lw/on/reductionism/*
7. *http://lesswrong.com/lw/uk/beyond_the_reach_of_god/*
8. *http://lesswrong.com/lw/m8/the_amazing_virgin_pregnancy/*

evolutionary biology can see the design signature of natural selection[9] on every wolf that chases a rabbit; and every rabbit that runs away; and every bee that stings instead of broadcasting a polite warning—but the deceptions of creationists sound plausible to *them,* I'm sure.

Not all lies are uncovered, not all liars are punished; we don't live in that righteous a universe. But not all lies are as safe as their liars believe. How many sins would become known to a Bayesian superintelligence, I wonder, if it did a (non-destructive?) nanotechnological scan of the Earth? At minimum, all the lies of which any evidence still exists in any brain. Some such lies may become known sooner than that, if the neuroscientists ever succeed in building a really good lie detector via neuroimaging. Paul Ekman (a pioneer in the study of tiny facial muscle movements) could probably read off a sizeable fraction of the world's lies right now, given a chance.

Not all lies are uncovered, not all liars are punished. But the Great Web is very commonly underestimated. Just the knowledge that humans have *already accumulated* would take many human lifetimes to learn[10]. Anyone who thinks that a non-God can tell a *perfect* lie, risk-free, is underestimating the tangledness of the Great Web.

Is honesty the best policy? I don't know if I'd go that far: Even on my ethics, it's sometimes okay to shut up. But compared to outright lies, either honesty or silence involves less exposure to recursively propagating risks you don't know you're taking.

9. *http://lesswrong.com/lw/kr/an_alien_god/*
10. *http://lesswrong.com/lw/kj/no_one_knows_what_science_doesnt_know/*

15. Of Lies and Black Swan Blowups[1]

Followup to: Entangled Truths, Contagious Lies[2]

Judge Marcus Einfeld, age 70, Queens Counsel since 1977, Australian Living Treasure 1997, United Nations Peace Award 2002, founding president of Australia's Human Rights and Equal Opportunities Commission, retired a few years back but routinely brought back to judge important cases...

...is going to jail for at least two years over a series of perjuries and lies that started with a £36, 6mph-over speeding ticket[3].

That whole *suspiciously virtuous-sounding* theory[4] about honest people not being good at lying, and entangled traces being left somewhere, and the entire thing blowing up in a Black Swan epic fail, actually *does* have a certain number of exemplars in real life, though obvious selective reporting is at work in our hearing about this one.

1. *http://lesswrong.com/lw/9a/of_lies_and_black_swan_blowups/*
2. Page 380, 'Entangled Truths, Contagious Lies'.
3. *http://news.bbc.co.uk/2/hi/uk_news/magazine/7967982.stm*
4. Page 380, 'Entangled Truths, Contagious Lies'.

16. Dark Side Epistemology[1]

Followup to: Entangled Truths, Contagious Lies[2]

If you once tell a lie, the truth is ever after your enemy.

I have previously spoken[3] of the notion that, the truth being entangled, lies are contagious. If you pick up a pebble from the driveway, and tell a geologist that you found it on a beach—well, do *you* know what a geologist knows about rocks? I don't. But I can suspect that a water-worn pebble wouldn't look like a droplet of frozen lava from a volcanic eruption. Do you know where the pebble in your driveway really came from? Things bear the marks of their places in a lawful universe; in that web, a lie is out of place. [Edit: Geologist in comments says that most pebbles in driveways are taken *from* beaches, so they couldn't tell the difference between a driveway pebble and a beach pebble, but they could tell the difference between a mountain pebble and a driveway/beach pebble. Case in point...]

What sounds like an arbitrary truth to one mind—one that could easily be replaced by a plausible lie—might be nailed down by a dozen linkages to the eyes of greater knowledge. To a creationist, the idea that life was shaped by "intelligent design" instead of "natural selection[4]" might sound like a sports team to cheer for. To a biologist, plausibly arguing that an organism was intelligently designed would require lying about almost every facet of the organism. To plausibly argue that "humans" were intelligently designed, you'd have to lie about the design of the human retina, the architecture of the human brain, the proteins bound together by weak van der Waals forces instead of strong covalent bonds...

1. *http://lesswrong.com/lw/uy/dark_side_epistemology/*
2. Page 380, 'Entangled Truths, Contagious Lies'.
3. Page 380, 'Entangled Truths, Contagious Lies'.
4. *http://lesswrong.com/lw/kr/an_alien_god/*

Or you could just lie about evolutionary theory, which is the path taken by most creationists. Instead of lying about the connected nodes in the network, they lie about the *general* laws governing the links.

And then to cover *that* up, they lie about the rules of science—like what it means to call something a "theory", or what it means for a scientist to say that they are not absolutely certain.

So they pass from lying about specific facts, to lying about general laws, to lying about the rules of reasoning. To lie about whether humans evolved, you must lie about evolution; and then you have to lie about the rules of science that constrain our understanding of evolution.

But how else? Just as a human would be out of place in a community of *actually* intelligently designed life forms, and you have to lie about the rules of evolution to make it appear otherwise; so too, beliefs about creationism are themselves out of place in science—you wouldn't find them in a well-ordered mind any more than you'd find palm trees growing on a glacier. And so you have to disrupt the barriers that would forbid them.

Which brings us to the case of self-deception.

A single lie you tell *yourself* may seem plausible enough, when you don't know any of the rules governing thoughts, or even that there *are* rules; and the choice[5] seems as arbitrary as choosing a flavor of ice cream, as isolated as a pebble on the shore...

...but then someone calls you on your belief, using the rules of reasoning that *they've* learned. They say, "Where's your evidence?"

And you say, "What? Why do I need evidence?"

So they say, "In general, beliefs require evidence."

This argument, clearly, is a soldier fighting on the other side[6], which you must defeat. So you say: "I disagree! Not all

5. *http://lesswrong.com/lw/rb/possibility_and_couldness/*
6. Page 148, 'Politics is the Mind-Killer'.

beliefs require evidence. In particular, beliefs about dragons don't require evidence. When it comes to dragons, you're allowed to believe anything you like. So I don't need evidence to believe there's a dragon in my garage[7]."

And the one says, "Eh? You can't just exclude dragons like that. There's a reason for the rule that beliefs require evidence. To draw a correct map[8] of the city, you have to walk through the streets[9] and make lines on paper that correspond to what you see. That's not an arbitrary legal requirement—if you sit in your living room and draw lines on the paper at random, the map's going to be wrong. With extremely high probability[10]. That's as true of a map of a dragon as it is of anything."

So now *this,* the explanation of *why* beliefs require evidence, is *also* an opposing soldier. So you say: "Wrong with extremely high probability? Then there's still a chance, right?[11] I don't have to believe if it's not absolutely certain[12]."

Or maybe you even begin to suspect, yourself, that "beliefs require evidence". But this threatens a lie you hold precious; so you reject the dawn inside you, push the sun back under the horizon.

Or you've previously heard the proverb "beliefs require evidence", and it sounded wise enough, and you endorsed it in public. But it never quite occurred to you, until someone else brought it to your attention, that this provreb could *apply to* your belief that there's a dragon in your garage. So you think fast and say, "The dragon is in a separate magisterium[13]."

7. Page 43, 'Belief in Belief'.
8. *http://yudkowsky.net/bayes/truth.html*
9. *http://lesswrong.com/lw/o5/the_second_law_of_thermodynamics_and_engines_of/*
10. *http://lesswrong.com/lw/o6/perpetual_motion_beliefs/*
11. Page 435, 'But There's Still A Chance, Right?'.
12. Page 443, 'Absolute Authority'.
13. *http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/*

Having false beliefs isn't a good thing, but it doesn't have to be permanently crippling—if, when you discover your mistake, you get over it. The dangerous thing is to have a false belief that you *believe should be protected as a belief*—a belief-in-belief[14], whether or not accompanied by actual belief.

A single Lie That Must Be Protected can block someone's progress into advanced rationality. No, it's not harmless fun.

Just as the world itself is more tangled by far[15] than it appears on the surface; so too, there are stricter rules of reasoning, constraining belief more strongly, than the untrained would suspect. The world is woven tightly, governed by general laws, and so are *rational* beliefs.

Think of what it would take to deny evolution or heliocentrism—all the connected truths and governing laws you wouldn't be allowed to know. Then you can imagine how a single act of self-deception can block off the whole meta-level of truthseeking, once your mind begins to be threatened by seeing the connections. Forbidding all the intermediate and higher levels of the rationalist's Art. Creating, in its stead, a vast complex of anti-law, rules of anti-thought, general justifications for believing the untrue.

Steven Kaas said[16], "Promoting less than maximally accurate beliefs is an act of sabotage. Don't do it to anyone unless you'd also slash their tires." Giving someone a false belief *to protect*—convincing them that the *belief itself* must be defended from any thought that seems to threaten it—well, you shouldn't do that to someone unless you'd also give them a frontal lobotomy.

Once you tell a lie, the truth is your enemy; and every truth connected to that truth, and every ally of truth in general; all of these you must oppose, to protect the lie. Whether you're lying to others, or to yourself.

14. Page 43, 'Belief in Belief'.
15. Page 380, 'Entangled Truths, Contagious Lies'.
16. *http://www.acceleratingfuture.com/steven/?p=124*

You have to deny that beliefs require evidence, and then you have to deny that maps should reflect territories, and then you have to deny that truth is a good thing...

Thus comes into being the Dark Side.

I worry that people aren't aware of it, or aren't sufficiently wary—that as we wander through our human world, we can expect to encounter *systematically* bad epistemology.

The "how to think" memes floating around, the cached thoughts[17] of Deep Wisdom[18]—some of it will be good advice devised by rationalists. But other notions were invented to protect a lie or self-deception: spawned from the Dark Side.

"Everyone has a right to their own opinion." When you think about it, where was that proverb generated? Is it something that someone would say in the course of protecting a truth, or in the course of protecting *from* the truth? But people don't perk up and say, "Aha! I sense the presence of the Dark Side!" As far as I can tell, it's not widely realized that the Dark Side is out there.

But how else? Whether you're deceiving others, or just yourself, the Lie That Must Be Protected will propagate recursively through the network of empirical causality, and the network of general empirical rules, and the rules of reasoning themselves, and the understanding behind those rules. If there is *good* epistemology in the world, and also lies or self-deceptions that people are trying to protect, then there will come into existence bad epistemology to counter the good. We could hardly expect, in this world, to find the Light Side without the Dark Side; there is the Sun, and that which shrinks away and generates a cloaking Shadow.

Mind you, these are not necessarily *evil* people[19]. The vast majority who go about repeating the Deep Wisdom are more

17. Page 297, 'Cached Thoughts'.
18. Page 314, 'How to Seem (and Be) Deep'.
19. Page 160, 'Are Your Enemies Innately Evil?'.

duped than duplicitous, more self-deceived than deceiving. I think.

And it's surely not my intent to offer you a Fully General Counterargument[20], so that whenever someone offers you some epistemology you don't like, you say: "Oh, someone on the Dark Side made that up." It's one of the rules of the Light Side that you have to refute the proposition for itself, not by accusing its inventor of bad intentions[21].

But the Dark Side is out there. Fear is the path that leads to it, and one betrayal can turn you. Not all who wear robes are either Jedi or fakes; there are also the Sith Lords, masters and unwitting apprentices. Be warned, be wary.

As for listing common memes that were spawned by the Dark Side—not random false beliefs, mind you, but bad epistemology, the Generic Defenses of Fail—well, would you care to take a stab at it, dear readers?

20. Page 333, 'Knowing About Biases Can Hurt People'.
21. Page 327, 'The Genetic Fallacy'.

17. The Sacred Mundane[1]

Followup to: Is Humanism a Religion-Substitute?[2]

So I was reading (around the first half of) Adam Frank's *The Constant Fire,* in preparation for my Bloggingheads dialogue[3] with him. Adam Frank's book is about the experience of the sacred. I might not usually call it that, but of course I know the experience Frank is talking about. It's what I feel when I watch a video of a space shuttle launch; or what I feel—to a lesser extent, because in this world it is too common[4]—when I look up at the stars at night, and think about what they mean. Or the birth of a child, say. That which is significant in the Unfolding Story.

Adam Frank holds that this experience is something that science holds deeply in common with religion. As opposed to e.g. being a basic human quality which religion corrupts.

The Constant Fire quotes William James's *The Varieties of Religious Experience* as saying:

> Religion... shall mean for us the feelings, acts, and experiences of individual men in their solitude; so far as they apprehend themselves to stand in relation to whatever they may consider the divine.

And this theme is developed further: Sacredness is something intensely *private* and *individual.*

Which completely nonplussed me. Am I supposed to not have any feeling of sacredness if I'm one of *many* people watching the video of *SpaceShipOne* winning the X-Prize? Why not? Am I supposed to think that my experience of sacredness has to

1. *http://lesswrong.com/lw/57/the_sacred_mundane/*
2. *http://lesswrong.com/lw/oy/is_humanism_a_religionsubstitute/*
3. *http://lesswrong.com/lw/4i/bhtv_yudkowsky_adam_frank_on_religious_experience/*
4. *http://lesswrong.com/lw/oz/scarcity/*

be somehow *different* from that of all the *other* people watching? Why, when we all have the same brain design[5]? Indeed, why would I *need* to believe I was unique? (But "unique" *is* another word Adam Frank uses; so-and-so's "unique experience of the sacred".) Is the feeling private in the same sense that we have difficulty communicating *any* experience? Then why emphasize this of sacredness, rather than sneezing?

The light came on when I realized that I was looking at a trick of Dark Side Epistemology[6]—if you make something *private,* that shields it from criticism. You can say, "You can't criticize me, because this is my private, inner experience that you can never access to question it."

But the price of shielding yourself from criticism is that you are cast into solitude—the solitude that William James admired as the core of religious experience, as if loneliness were a *good* thing.

Such relics of Dark Side Epistemology are key to understanding the many ways that religion twists the experience of sacredness:

Mysteriousness—why should the sacred have to be mysterious? A space shuttle launch gets by just fine without being mysterious. How much *less* would I appreciate the stars if I did *not* know what they were, if they were just little points in the night sky? But if your religious beliefs are questioned—if someone asks, "Why doesn't God heal amputees?"—then you take refuge and say, in a tone of deep profundity, "It is a sacred mystery!" There are questions that must not be asked, and answers that must not be acknowledged, to defend the lie. Thus unanswerability comes to be associated with sacredness. And the price of shielding yourself from criticism is giving up the true curiosity[7] that truly wishes to find answers. You will worship your own ignorance of the temporarily unanswered questions

5. *http://lesswrong.com/lw/rl/the_psychological_unity_of_humankind/*
6. Page 385, 'Dark Side Epistemology'.
7. Page 478, 'The Meditation on Curiosity'.

of your own generation—probably including[8] ones that are already answered[9].

Faith—in the early days of religion, when people were more naive, when even intelligent folk actually believed that stuff, religions staked their reputation upon the testimony of miracles in their scriptures. And Christian archaeologists set forth truly expecting to find the ruins of Noah's Ark. But when no such evidence was forthcoming, *then* religion executed what William Bartley called *the retreat to commitment,* "I believe because I believe!" Thus *belief without good evidence* came to be associated with the experience of the sacred. And the price of shielding yourself from criticism is that you sacrifice your ability to think clearly about that which is sacred, and to progress in your understanding of the sacred, and relinquish mistakes.

Experientialism—if before you thought that the rainbow was a sacred contract of God with humanity, and then you begin to realize that God doesn't exist, then you may execute a *retreat to pure experience*—to praise yourself just for *feeling* such wonderful sensations when you think about God, whether or not God actually *exists*. And the price of shielding yourself from criticism is solipsism: your experience is stripped of its *referents*. What a terrible hollow feeling it would be to watch a space shuttle rising on a pillar of flame, and say to yourself, "But it doesn't really matter whether the space shuttle actually exists, so long as I feel."

Separation—if the sacred realm is not subject to ordinary rules of evidence or investigable by ordinary means, then it must be different in kind from the world of mundane matter: and so we are less likely to think of a space shuttle as a candidate for sacredness, because it is a work of merely *human* hands. Keats lost his admiration of the rainbow[10] and demoted it to the "dull catalogue of mundane things" for the crime of its

8. *http://lesswrong.com/lw/kj/no_one_knows_what_science_doesnt_know/*
9. *http://lesswrong.com/lw/r5/the_quantum_physics_sequence/*
10. *http://lesswrong.com/lw/or/joy_in_the_merely_real/*

woof and texture being known. And the price of shielding yourself from all ordinary criticism is that you lose the sacredness of all merely real[11] things.

Privacy—of this I have already spoken.

Such distortions are why we had best *not* to try to salvage religion. No, not even in the form of "spirituality". Take away the institutions and the factual mistakes, subtract the churches and the scriptures, and you're left with... all this nonsense about mysteriousness, faith, solipsistic experience, private solitude, and discontinuity.

The original lie is only the beginning of the problem. Then you have all the ill habits of thought that have evolved to defend it. Religion is a poisoned chalice, from which we had best not even sip. Spirituality is the same cup after the original pellet of poison has been taken out, and only the dissolved portion remains—a little less directly lethal, but still not good for you.

When a lie has been defended for ages upon ages, the true origin of the inherited habits lost in the mists, with layer after layer of undocumented sickness; then the wise, I think, will start over from scratch, rather than trying to selectively discard the original lie while keeping the habits of thought that protected it. *Just admit you were wrong,* give up *entirely* on the mistake, stop defending it *at all*, stop trying to say you were even a little right, stop trying to save face, just say "Oops![12]" and throw out the *whole* thing and begin again.

That capacity—to really, *really,* without defense, admit you were *entirely* wrong—is why religious experience will never be like scientific experience. No religion can absorb *that* capacity without losing itself *entirely* and becoming simple humanity...

...to just look up at the distant stars. Believable without strain, without a constant distracting struggle to fend off your awareness of the counterevidence. Truly there *in the world,* the experience united with the referent, a solid part of that un-

11. *http://lesswrong.com/lw/or/joy_in_the_merely_real/*
12. Page 466, 'The Importance of Saying "Oops"'.

folding story. Knowable without threat, offering true meat for curiosity. Shared in togetherness with the many other onlookers, no need to retreat to privacy. Made of the same fabric as yourself and all other things. Most holy and beautiful, the sacred mundane.

Against Doublethink

1. Singlethink[1]

I remember the exact moment when I began my journey as a rationalist.

It was not while reading *Surely You're Joking, Mr. Feynman* or any existing work upon rationality; for these I simply accepted as obvious. The journey begins when you see a great flaw in your existing art, and discover a drive to improve, to create *new* skills beyond the helpful but inadequate ones you found in books.

In the last moments of my first life, I was fifteen years old, and rehearsing a pleasantly self-righteous memory of a time when I was much younger. My memories this far back are vague; I have a mental image, but I don't remember how old I was exactly. I think I was six or seven, and that the original event happened during summer camp.

What happened originally was that a camp counselor, a teenage male, got us much younger boys to form a line, and proposed the following game: the boy at the end of the line would crawl through our legs, and we would spank him as he went past, and then it would be the turn of the next eight-year-old boy at the end of the line. (Maybe it's just that I've lost my youthful innocence, but I can't help but wonder...) I refused to play this game, and was told to go sit in the corner.

This memory—of refusing to spank and be spanked—came to symbolize to me that even at this very early age I had refused to take joy in hurting others. That I would not purchase a spank on another's butt, at the price of a spank on my own; would not pay in hurt for the opportunity to inflict hurt. I had refused to play a negative-sum game.

And then, at the age of fifteen, I suddenly realized that it wasn't true. I *hadn't* refused out of a principled stand against negative-sum games. I found out about the Prisoner's Dilemma pretty early in life, but not at the age of seven. I'd refused sim-

1. *http://lesswrong.com/lw/ko/singlethink/*

ply because I didn't want to get hurt, and standing in the corner was an acceptable price to pay for not getting hurt.

More importantly, I realized that I had *always* known this—that the real memory had *always* been lurking in a corner of my mind, my mental eye glancing at it for a fraction of a second and then looking away.

In my very first step along the Way, *I caught the feeling*—generalized over the subjective experience—and said, "So *that's* what it feels like to shove an unwanted truth into the corner of my mind! Now I'm going to notice every time I do that, and clean out *all* my corners!"

This discipline I named *singlethink,* after Orwell's doublethink. In doublethink,[2] you forget, and then forget you have forgotten. In singlethink, you notice you are forgetting, and then you remember. You hold only a single non-contradictory thought in your mind at once.

"Singlethink" was the first *new* rationalist skill I created, which I had not read about in books. I doubt that it is original in the sense of academic priority, but this is thankfully not required.

Oh, and my fifteen-year-old self liked to name things.

The terrifying depths of the confirmation bias go on and on. Not forever, for the brain is of finite complexity, but long enough that it feels like forever. You keep on discovering (or reading about) new mechanisms by which your brain shoves things out of the way.

But my young self swept out quite a few corners with that first broom.

2. Page 401, 'Doublethink (Choosing to be Biased)'.

2. Doublethink (Choosing to be Biased)[1]

An oblong slip of newspaper had appeared between O'Brien's fingers. For perhaps five seconds it was within the angle of Winston's vision. It was a photograph, and there was no question of its identity. It was the photograph. It was another copy of the photograph of Jones, Aaronson, and Rutherford at the party function in New York, which he had chanced upon eleven years ago and promptly destroyed. For only an instant it was before his eyes, then it was out of sight again. But he had seen it, unquestionably he had seen it! He made a desperate, agonizing effort to wrench the top half of his body free. It was impossible to move so much as a centimetre in any direction. For the moment he had even forgotten the dial. All he wanted was to hold the photograph in his fingers again, or at least to see it.

'It exists!' he cried.

'No,' said O'Brien.

He stepped across the room.

There was a memory hole in the opposite wall. O'Brien lifted the grating. Unseen, the frail slip of paper was whirling away on the current of warm air; it was vanishing in a flash of flame. O'Brien turned away from the wall.

'Ashes,' he said. 'Not even identifiable ashes. Dust. It does not exist. It never existed.'

1. *http://lesswrong.com/lw/je/doublethink_choosing_to_be_biased/*

> 'But it did exist! It does exist! It exists in memory. I remember it. You remember it.'
>
> 'I do not remember it,' said O'Brien.
>
> Winston's heart sank. That was doublethink. He had a feeling of deadly helplessness. If he could have been certain that O'Brien was lying, it would not have seemed to matter. But it was perfectly possible that O'Brien had really forgotten the photograph. And if so, then already he would have forgotten his denial of remembering it, and forgotten the act of forgetting. How could one be sure that it was simple trickery? Perhaps that lunatic dislocation in the mind could really happen: that was the thought that defeated him.
>
> —George Orwell[2], *1984*

What if self-deception helps us be happy? What if just running out and overcoming bias will make us—gasp!—*unhappy?* Surely, *true* wisdom would be *second-order* rationality, choosing when to be rational. That way you can decide which cognitive biases should govern you, to maximize your happiness.

Leaving the morality aside, I doubt such a lunatic dislocation in the mind could really happen.

Second-order rationality implies that at some point, you will think to yourself, "And now, I will irrationally believe that I will win the lottery, in order to make myself happy." But we do not have such direct control over our beliefs. You cannot make yourself believe the sky is green by an act of will. You might be able to believe you believed[3] it—though I have just made that more difficult for you by pointing out the difference. (You're

2. Page 191, 'Human Evil and Muddled Thinking'.
3. Page 43, 'Belief in Belief'.

welcome!) You might even *believe* you were happy and self-deceived; but you would not *in fact* be happy and self-deceived.

For second-order rationality to be genuinely *rational,* you would first need a good model of reality, to extrapolate the consequences of rationality and irrationality. If you then chose to be first-order irrational, you would need to forget this accurate view. And then forget the act of forgetting. I don't mean to commit the logical fallacy of generalizing from fictional evidence, but I think Orwell did a good job of extrapolating where this path leads.

You can't know the consequences of being biased, until you have already debiased yourself. And then it is too late for self-deception.

The other alternative is to choose blindly to remain biased, without any clear idea of the consequences. This is not second-order rationality. It is willful stupidity.

Be irrationally optimistic about your driving skills, and you will be happily unconcerned where others sweat and fear. You won't have to put up with the inconvenience of a seatbelt. You will be happily unconcerned for a day, a week, a year. Then *CRASH,* and spend the rest of your life wishing you could scratch the itch in your phantom limb. Or paralyzed from the neck down. Or dead. It's not inevitable, but it's possible; how probable is it? You can't make that tradeoff rationally unless you know your *real* driving skills, so you can figure out how much danger you're placing yourself in. You can't make that tradeoff rationally unless you know about biases like neglect of probability[4].

No matter how many days go by in blissful ignorance, it only takes a single mistake to undo a human life, to outweigh every penny you picked up from the railroad tracks of stupidity.

One of chief pieces of advice I give to aspiring rationalists is "Don't try to be clever." And, "Listen to those quiet, nagging doubts." If you don't know, you don't know *what* you don't

4. *http://en.wikipedia.org/wiki/Neglect_of_probability*

know, you don't know how *much* you don't know, and you don't know how much you *needed* to know.

There is no second-order rationality. There is only a blind leap into what may or may not be a flaming lava pit. Once you *know*, it will be too late for blindness.

But people neglect this, because they do not know what they do not know. Unknown unknowns are not available[5]. They do not focus on the blank area on the map, but treat it as if it corresponded to a blank territory. When they consider leaping blindly, they check their memory for dangers, and find no flaming lava pits in the blank map. Why not leap?

Been there. Tried that. Got burned. Don't try to be clever.

I once said to a friend that I suspected the happiness of stupidity was greatly overrated. And she shook her head seriously, and said, "No, it's not; it's really not."

Maybe there are stupid happy people out there. Maybe they are happier than you are. And life isn't fair, and you won't become happier by being jealous of what you can't have. I suspect the vast majority of *Overcoming Bias* readers could not achieve the "happiness of stupidity" if they tried. That way is closed to you. You can never achieve that degree of ignorance, you cannot forget what you know, you cannot unsee what you see.

The happiness of stupidity is closed to you. You will never have it short of actual brain damage, and maybe not even then. You should wonder, I think, whether the happiness of stupidity is *optimal*—if it is the *most* happiness that a human can aspire to—but it matters not. That way is closed to you, if it was ever open.

All that is left to you now, is to aspire to such happiness as a rationalist can achieve. I think it may prove greater, in the end. There are bounded paths and open-ended paths; plateaus on which to laze, and mountains to climb; and if climbing takes more effort, still the mountain rises higher in the end.

5. *http://lesswrong.com/lw/j5/availability/*

Also there is more to life than happiness; and other happinesses than your own may be at stake in your decisions.

But that is moot. By the time you realize you have a choice, there is no choice. You cannot unsee what you see. The other way is closed.

3. No, Really, I've Deceived Myself[1]

Followup to: Belief in Belief[2]
3

I recently spoke with a person who... it's difficult to describe. Nominally, she was an Orthodox Jew. She was also highly intelligent, conversant with some of the archaeological evidence against her religion, and the shallow standard arguments against religion that religious people know about. For example, she knew that Mordecai, Esther, Haman, and Vashti were not in the Persian historical records, but that there was a corresponding old Persian legend about the Babylonian gods Marduk and Ishtar, and the rival Elamite gods Humman and Vashti. She *knows* this, and she still celebrates Purim. One of those highly intelligent religious people who stew in their own contradictions for years, elaborating and tweaking, until their minds look like the inside of an M. C. Escher painting.

Most people like this will pretend that they are much too wise[4] to talk to atheists, but she was willing to talk with me for a few hours.

As a result, I now understand at least one more thing about self-deception that I didn't explicitly understand before—namely, that you don't have to *really* deceive yourself so long as you *believe* you've deceived yourself. Call it "belief in self-deception".

When this woman was in high school, she thought she was an atheist. But she decided, at that time, that she should act as if she believed in God. And then—she told me earnestly—over time, she came to really believe in God.

So far as I can tell, she is completely wrong about that. Always throughout our conversation, she said, over and over, "I

1. *http://lesswrong.com/lw/r/no_really_ive_deceived_myself/*
2. Page 43, 'Belief in Belief'.
3. Page 43, 'Belief in Belief'.
4. *http://lesswrong.com/lw/yp/pretending_to_be_wise/*

believe in God", never once, "There *is* a God." When I asked her why she was religious, she never once talked about the consequences of God existing, only about the consequences of believing in God. Never, "God will help me", always, "my belief in God helps me". When I put to her, "Someone who just wanted the truth and looked at our universe would not even invent God as a hypothesis," she agreed outright.

She hasn't *actually* deceived herself into believing that God exists or that the Jewish religion is true. Not even close, so far as I can tell.

On the other hand, I think she really *does* believe she has deceived herself.

So although she does not receive any benefit of believing in God—because she doesn't—she honestly *believes* she has deceived herself into believing in God, and so she honestly *expects* to receive the benefits that she associates with deceiving oneself into believing in God; and *that,* I suppose, ought to produce much the same placebo effect as *actually* believing in God.

And this may explain why she was motivated to earnestly defend the statement that she *believed* in God from my skeptical questioning, while never saying "Oh, and by the way, God actually does exist" or even seeming the slightest bit interested in the proposition.

4. Belief in Self-Deception[1]

Continuation of: No, Really, I've Deceived Myself[2]

Followup to: Dark Side Epistemology[3]

I spoke yesterday of my conversation with a nominally Orthodox Jewish woman who vigorously defended the assertion that she believed in God, while seeming not to actually believe in God at all.

While I was questioning her about the benefits that she thought came from believing in God, I introduced the Litany of Tarski[4]—which is actually an infinite family of litanies, a specific example being:

If the sky is blue
I desire to believe "the sky is blue"
If the sky is not blue
I desire to believe "the sky is not blue".

"This is not my philosophy," she said to me.

"I didn't think it was," I replied to her. "I'm just asking—assuming that God does *not* exist, and this is known, then should you still believe in God?"

She hesitated. She seemed to really be trying to think about it, which surprised me.

"So it's a counterfactual question..." she said slowly.

I thought at the time that she was having difficulty allowing herself to visualize the world where God does not exist, because of her attachment to a God-containing world.

Now, however, I suspect she was having difficulty visualizing a contrast between the way the *world* would look if God existed or did not exist, because all her thoughts were about her *belief in God*, but her causal network modelling the world did

1. *http://lesswrong.com/lw/s/belief_in_selfdeception/*
2. Page 406, 'No, Really, I've Deceived Myself'.
3. Page 43, 'Belief in Belief'.
4. Page 478, 'The Meditation on Curiosity'.

not contain God as a node. So she could easily answer "How would the world look different if I didn't believe in God?", but not "How would the world look different if there was no God?"

She didn't answer that question, at the time. But she did produce a *counterexample* to the Litany of Tarski:

She said, "I believe that people are nicer than they really are."

I tried to explain that if you say, "People are bad," that means you believe people are bad, and if you say, "I believe people are nice", that means you believe you believe people are nice. So saying "People are bad and I believe people are nice" means you believe people are bad but you believe you believe people are nice.

I quoted to her:

> "If there were a verb meaning 'to believe falsely', it would not have any
> significant first person, present indicative."
> —Ludwig Wittgenstein

She said, smiling, "Yes, I believe people are nicer than, in fact, they are. I just thought I should put it that way for you."

> "I reckon Granny ought to have a good look at you, Walter," said Nanny. "I reckon
> your mind's all tangled up like a ball of string what's been dropped."
> —Terry Pratchett, *Maskerade*

And I can type out the words, "Well, I guess she didn't believe that her reasoning ought to be consistent under reflection[5]," but I'm still having trouble coming to grips[6] with it.

I can see the pattern in the words coming out of her lips, but I can't understand the mind behind on an empathic level. I can imagine myself into the shoes of baby-eating aliens[7] and the La-

5. Page 34, 'The Lens That Sees Its Flaws'.
6. *http://lesswrong.com/lw/hs/think_like_reality/*
7. *http://lesswrong.com/lw/y5/the_babyeating_aliens_18/*

dy 3rd Kiritsugu[8], but I cannot imagine what it is like to be her. Or maybe I just don't *want* to?

This is why intelligent people only have a certain amount of time (measured in subjective time spent thinking about religion) to become atheists. After a certain point, if you're smart, have spent time thinking about and defending your religion, and still haven't escaped the grip of Dark Side Epistemology[9], the inside of your mind ends up as an Escher painting.

(One of the other few moments that gave her pause—I mention this, in case you have occasion to use it—is when she was talking about how it's good to believe that someone cares whether you do right or wrong—*not*, of course, talking about how there actually *is* a God who cares whether you do right or wrong, this proposition is not part of her religion—

And I said, "But *I* care whether you do right or wrong. So what you're saying is that this isn't enough, and you also need to believe in something *above* humanity that cares whether you do right or wrong." So that stopped her, for a bit, because of course she'd never thought of it in those terms before. Just a standard application of the nonstandard toolbox[10].)

Later on, at one point, I was asking her if it would be good to do *anything* differently if there definitely was no God, and this time, she answered, "No."

"So," I said incredulously, "if God exists or doesn't exist, that has absolutely no effect on how it would be good for people to think or act? I think even a rabbi would look a little askance at that."

Her religion seems to now consist *entirely* of the worship of worship. As the true believers of older times might have believed that an all-seeing father would save them, she now believes that belief in God will save her.

8. *http://lesswrong.com/lw/y7/the_super_happy_people_38/*
9. Page 385, 'Dark Side Epistemology'.
10. Page 314, 'How to Seem (and Be) Deep'.

After she said "I believe people are nicer than they are," I asked, "So, are you consistently surprised when people undershoot your expectations?" There was a long silence, and then, slowly: "Well... am I *surprised* when people... undershoot my expectations?"

I didn't understand this pause at the time. I'd intended it to suggest that if she was constantly disappointed by reality, then this was a downside of believing falsely. But she seemed, instead, to be taken aback at the implications of *not* being surprised.

I now realize that the whole essence of her philosophy was *her belief that she had deceived herself,* and the possibility that her estimates of other people were *actually accurate*, threatened the Dark Side Epistemology[11] that she had built around beliefs such as "I benefit from believing people are nicer than they actually are."

She has taken the old idol off its throne, and replaced it with an explicit worship of the Dark Side Epistemology that was once invented to defend the idol; she worships her own attempt at self-deception. The attempt failed, but she is honestly unaware of this.

And so humanity's token guardians of sanity (motto: "pooping your deranged little party since Epicurus") must now fight the active worship of self-deception—the worship *of the supposed benefits of faith*, in place of God.

This actually explains a fact about *myself* that I didn't really understand earlier—the reason why I'm annoyed when people talk as if self-deception is *easy,* and why I write entire blog posts[12] arguing that making a deliberate choice to believe the sky is green, is harder to get away with than people seem to think.

It's because—while you *can't* just choose to believe the sky is green—if you don't *realize* this fact, then you actually *can* fool

11. Page 385, 'Dark Side Epistemology'.
12. Page 401, 'Doublethink (Choosing to be Biased)'.

yourself into believing that you've successfully deceived yourself.

And since you then sincerely *expect* to receive the benefits that you think come from self-deception, you get the same sort of placebo benefit that would actually come from a successful self-deception.

So by going around explaining how *hard* self-deception is, I'm actually taking direct aim at the placebo benefits that people get from believing that they've deceived themselves, and targeting the new sort of religion that worships only the worship of God.

Will this battle, I wonder, generate a new list of reasons why, not belief, but belief in belief[13], is *itself* a good thing? Why people derive great benefits from worshipping their worship? Will we have to do this over again with belief in belief in belief and worship of worship of worship? Or will intelligent theists finally just give up on that line of argument?

I wish I could believe that no one could possibly believe in belief in belief in belief, but the Zombie World argument in philosophy has gotten even more tangled than this[14] and its proponents still haven't abandoned it.

I await the eager defenses of belief in belief in the comments, but I wonder if anyone would care to jump ahead of the game and defend belief in belief in belief? Might as well go ahead and get it over with.

13. Page 43, 'Belief in Belief'.
14. *http://lesswrong.com/lw/p7/zombies_zombies/*

5. Moore's Paradox[1]

Followup to: Belief in Self-Deception[2]

Moore's Paradox[3] is the standard term for saying "It's raining outside but I don't believe that it is." HT to painquale on MetaFilter.[4]

I think I understand Moore's Paradox a bit better now, after reading some of the comments on Less Wrong. Jimrandomh[5] suggests:

> Many people cannot distinguish between levels of indirection. To them, "I believe X" and "X" are the same thing, and therefore, reasons why it is beneficial to believe X are also reasons why X is true.

I don't think this is correct—relatively young children can understand the concept of having a false belief, which requires separate mental buckets for the map and the territory. But it points in the direction of a similar idea:

Many people may not consciously distinguish between *believing* something and *endorsing* it.

After all—"I believe in democracy" means, colloquially, that you endorse the concept of democracy, not that you believe democracy exists. The word "belief", then, has more than one meaning. We could be looking at a confused word[6] that causes confused thinking (or maybe it just reflects pre-existing confusion).

1. *http://lesswrong.com/lw/1f/moores_paradox/*
2. Page 408, 'Belief in Self-Deception'.
3. *http://en.wikipedia.org/wiki/Moore%27s_paradox*
4. *http://www.metafilter.com/79752/More-Right-was-too-political#2477702*
5. *http://lesswrong.com/lw/r/no_really_ive_deceived_myself/#ga*
6. *http://lesswrong.com/lw/nw/fallacies_of_compression/*

So: in the original example[7], "I believe people are nicer than they are", she came up with some reasons why it would be good to believe people are nice—health benefits and such—and since she now had some warm affect on "believing people are nice", she introspected on this warm affect and concluded, "I believe people are nice". That is, she mistook the *positive affect* attached to the quoted belief, as signaling *her belief in the proposition*. At the same time, the world itself seemed like people weren't so nice. So she said, "I believe people are nicer than they are."

And that verges on being an honest mistake—sort of—since people are not taught explicitly how to know when they believe something. As in the parable of the dragon in the garage[8]; the one who says "There is a dragon in my garage—but it's invisible", does not recognize his *anticipation* of seeing no dragon, as indicating that he possesses an (accurate) model with no dragon in it.

It's not as if people are *trained* to recognize when they believe something. It's not like they're ever taught in high school: "What it feels like to actually believe something—to have that statement in your belief pool—is that it just seems like the way the world *is*. You should recognize this feeling, which is actual (unquoted) belief, and distinguish it from having good feelings about a belief that you recognize as a belief (which means that it's in quote marks)."

This goes a long way toward making this real-life case of Moore's Paradox seem less alien[9], and providing another mechanism whereby people can be simultaneously right and wrong[10].

Likewise Kurige[11] who wrote:

7. Page 408, 'Belief in Self-Deception'.
8. Page 43, 'Belief in Belief'.
9. Page 408, 'Belief in Self-Deception'.
10. *http://lesswrong.com/lw/1d/simultaneously_right_and_wrong/*
11. *http://lesswrong.com/lw/r/no_really_ive_deceived_myself/#gk*

> I believe that there is a God—and that He has instilled a sense of right and wrong in us by which we are able to evaluate the world around us. I also believe a sense of morality has been evolutionarily programmed into us—a sense of morality that is most likely a result of the formation of meta-political coalitions in Bonobo communities a very, very long time ago. These two beliefs are not contradictory, but the complexity lies in reconciling the two.

I suspect, Kurige, you have decided that you have *reasons to endorse* the quoted belief that God has instilled a sense of right and wrong in us. And also that you have reasons to endorse the verdict of science. They both seem like good communities to join, right? There are benefits to both sets of beliefs? You introspect and find that you feel good about both beliefs?

But you did *not* say:

"God instilled a sense of right and wrong in us, and also a sense of morality has been evolutionarily programmed into us. The two states of reality are not inconsistent, but the complexity lies in reconciling the two."

If you're reading this, Kurige, you should very quickly say the above out loud, so you can notice that it seems at least slightly harder to swallow—notice the *subjective difference*—before you go to the trouble of rerationalizing.

This is the subjective difference between having reasons to endorse two different beliefs, and your mental model of a single world, a single way-things-are.

6. Don't Believe You'll Self-Deceive[1]

Followup to: Moore's Paradox[2], Doublethink[3]

I don't mean to seem like I'm picking on Kurige, but I think you have to expect a certain amount of questioning if you show up on Less Wrong and say[4]:

> One thing I've come to realize that helps to explain the disparity I feel when I talk with most other Christians is the fact that somewhere along the way my world-view took a major shift away from blind faith and landed somewhere in the vicinity of Orwellian double-think.

"If you *know* it's double-think[5]...
...how can you still *believe* it?" I helplessly want to say.
Or[6]:

> I chose to believe in the existence of God—deliberately and consciously. This decision, however, has absolutely zero effect on the actual existence of God.

If you *know* your belief isn't correlated to reality, how can you still believe it?

Shouldn't the *gut-level* realization, "Oh, wait, the sky really *isn't* green" follow from the realization "My map that says 'the sky is green' has no reason to be correlated with the territory"?

1. *http://lesswrong.com/lw/1o/dont_believe_youll_selfdeceive/*
2. Page 413, 'Moore's Paradox'.
3. Page 413, 'Moore's Paradox'.
4. *http://lesswrong.com/lw/r/no_really_ive_deceived_myself/#gk*
5. Page 401, 'Doublethink (Choosing to be Biased)'.
6. *http://lesswrong.com/lw/1f/moores_paradox/#u3*

Well... apparently not.

One part of this puzzle may be my explanation of Moore's Paradox[7] ("It's raining, but I don't believe it is")—that people introspectively mistake positive affect attached to a quoted belief, for actual credulity.

But another part of it may just be that—contrary to the indignation I initially wanted to put forward—it's actually quite *easy* not to make the jump from "The map that reflects the territory would say 'X'" to actually believing "X". It takes some work to *explain* the ideas of minds as map-territory correspondence builders[8], and even then, it may take more work to get the implications on a *gut level*.

I realize now that when I wrote[9] "You cannot make yourself believe the sky is green by an act of will", I wasn't just a dispassionate reporter of the existing facts. I was also trying to instill a self-fulfilling prophecy.

It may be wise to go around deliberately repeating "I can't get away with double-thinking! Deep down, I'll know it's not true! If I know my map has no reason to be correlated with the territory, that means I don't believe it!"

Because that way—if you're ever tempted to try—the thoughts "But I know this isn't really true!" and "I can't fool myself!" will always rise readily to mind; and that way, you will indeed be less likely to fool yourself successfully. You're more likely to get, on a gut level, that telling yourself X doesn't make X true: and therefore, really truly not-X.

If you keep telling yourself that you *can't* just deliberately choose to believe the sky is green—then you're less likely to succeed in fooling yourself on one level or another; either in the sense of really believing it, or of falling into Moore's Paradox[10], belief in belief[11], or belief in self-deception[12].

7. Page 413, 'Moore's Paradox'.
8. Page 18, 'What is Evidence?'.
9. Page 413, 'Moore's Paradox'.
10. Page 413, 'Moore's Paradox'.

If you keep telling yourself that deep down you'll know—

If you keep telling yourself that you'd just look at your elaborately constructed false map, and just know that it was a false map without any expected correlation to the territory, and therefore, despite all its elaborate construction, you wouldn't be able to invest any credulity in it—

If you keep telling yourself that reflective consistency will take over and make you stop believing on the object level, once you come to the meta-level realization that the map is not reflecting—

Then when push comes to shove—you may, indeed, fail.

When it comes to deliberate self-deception, you must *believe in your own inability!*

Tell yourself the effort is doomed—*and it will be!*

Is that the power of positive thinking, or the power of negative thinking? Either way, it seems like a wise precaution.

11. Page 43, 'Belief in Belief'.
12. Page 408, 'Belief in Self-Deception'.

Overly Convenient Excuses

1. The Proper Use of Humility[1]

It is widely recognized that good science requires some kind of humility. *What sort* of humility is more controversial.

Consider the creationist who says: "But who can really know whether evolution is correct? It is just a theory. You should be more humble and open-minded." Is this humility? The creationist practices a very selective underconfidence, refusing to integrate massive weights of evidence in favor of a conclusion he finds uncomfortable. I would say that whether you call this "humility" or not, it is the wrong step in the dance.

What about the engineer who humbly designs fail-safe mechanisms into machinery, even though he's damn sure the machinery won't fail? This seems like a good kind of humility to me. Historically, it's not unheard-of for an engineer to be damn sure a new machine won't fail, and then it fails anyway.

What about the student who humbly double-checks the answers on his math test? Again I'd categorize that as good humility.

What about a student who says, "Well, no matter how many times I check, I can't ever be *certain* my test answers are correct," and therefore doesn't check even once? Even if this choice stems from an emotion similar to the emotion felt by the previous student, it is less wise.

You suggest studying harder, and the student replies: "No, it wouldn't work for me; I'm not one of the smart kids like you; nay, one so lowly as myself can hope for no better lot." This is social modesty, not humility. It has to do with regulating status in the tribe, rather than scientific process. If you ask someone to "be more humble", by default they'll associate the words to social modesty—which is an intuitive, everyday, ancestrally relevant concept. Scientific humility is a more recent and rarefied invention, and it is not inherently social. Scientific humility is something you would practice even if you were alone in a space-

1. *http://lesswrong.com/lw/gq/the_proper_use_of_humility/*

suit, light years from Earth with no one watching. Or even if you received an absolute guarantee that no one would ever criticize you again, no matter what you said or thought of yourself. You'd still double-check your calculations if you were wise.

The student says: "But I've seen other students double-check their answers and then they still turned out to be wrong. Or what if, by the problem of induction, 2 + 2 = 5 this time around? No matter what I do, I won't be sure of myself." It sounds very profound, and very modest. But it is not coincidence that the student wants to hand in the test quickly, and go home and play video games.

The end of an era in physics does not always announce itself with thunder and trumpets; more often it begins with what seems like a small, small flaw... But because physicists have this arrogant idea that their models should work *all* the time, not just *most* of the time, they follow up on small flaws. Usually, the small flaw goes away under closer inspection. Rarely, the flaw widens to the point where it blows up the whole theory. Therefore it is written: "If you do not seek perfection you will halt before taking your first steps."

But think of the social audacity of trying to be right *all* the time! I seriously suspect that if Science claimed that evolutionary theory is true most of the time but not all of the time—or if Science conceded that maybe on some days the Earth *is* flat, but who really knows—then scientists would have better social reputations. Science would be viewed as less confrontational, because we wouldn't have to argue with people who say the Earth is flat—there would be room for compromise. When you argue a lot, people look upon you as confrontational. If you repeatedly refuse to compromise, it's even worse. Consider it as a question of tribal status: scientists have certainly earned some extra status in exchange for such socially useful tools as medicine and cellphones. But this social status does not justify their insistence that *only* scientific ideas on evolution be taught in public schools. Priests also have high social status, after all.

Scientists are getting above themselves—they won a little status, and now they think they're chiefs of the whole tribe! They ought to be more humble, and compromise a little.

Many people seem to possess rather hazy views of "rationalist humility". It is dangerous to have a prescriptive principle which you only vaguely comprehend; your mental picture may have so many degrees of freedom that it can adapt to justify almost any deed. Where people have vague mental models that can be used to argue anything, they usually end up believing whatever they started out wanting to believe. This is so convenient that people are often reluctant to give up vagueness. But the purpose of our ethics is to move us, not be moved by us.

"Humility" is a virtue that is often misunderstood. This doesn't mean we should discard the concept of humility, but we should be careful using it. It may help to look at the *actions* recommended by a "humble" line of thinking, and ask: "Does acting this way make you stronger, or weaker?" If you think about the problem of induction as applied to a bridge that needs to stay up, it may sound reasonable to conclude that nothing is certain no matter what precautions are employed; but if you consider the real-world difference between adding a few extra cables, and shrugging, it seems clear enough what makes the stronger bridge.

The vast majority of appeals that I witness to "rationalist's humility" are excuses to shrug. The one who buys a lottery ticket, saying, "But you can't *know* that I'll lose." The one who disbelieves in evolution, saying, "But you can't *prove* to me that it's true." The one who refuses to confront a difficult-looking problem, saying, "It's probably too hard to solve." The problem is motivated skepticism[2] aka disconfirmation bias—more heavily scrutinizing assertions that we don't want to believe. Humility, in its most commonly misunderstood form, is a fully general excuse not to believe something; since, after all, you can't be *sure*. Beware of fully general excuses!

2. *http://www.sunysb.edu/polsci/mlodge/taber_lodge_ajps.pdf*

A further problem is that humility is all too easy to *profess*. Dennett, in "Breaking the Spell", points out that while many religious assertions are very hard to believe, it is easy for people to believe that they *ought* to believe them. Dennett terms this "belief in belief". What would it mean to really assume, to really believe, that three is equal to one? It's a lot easier to believe that you *should*, somehow, believe that three equals one, and to make this response at the appropriate points in church. Dennett suggests that much "religious belief" should be studied as "religious profession"—what people think they should believe and what they know they ought to say.

It is all too easy to meet every counterargument by saying, "Well, of course I could be wrong." Then, having dutifully genuflected in the direction of Modesty, having made the required obeisance, you can go on about your way without changing a thing.

The temptation is always to claim the most points with the least effort. The temptation is to carefully integrate all incoming news in a way that lets us change our beliefs, and above all our *actions*, as little as possible. John Kenneth Galbraith said: "Faced with the choice of changing one's mind and proving that there is no need to do so, almost everyone gets busy on the proof." And the greater the *inconvenience* of changing one's mind, the more effort people will expend on the proof.

But y'know, if you're gonna *do* the same thing anyway, there's no point in going to such incredible lengths to rationalize it. Often I have witnessed people encountering new information, apparently accepting it, and then carefully explaining why they are going to do exactly the same thing they planned to do previously, but with a different justification. The point of thinking is to *shape* our plans; if you're going to keep the same plans anyway, why bother going to all that work to justify it? When you encounter new information, the hard part is to *update*, to *react*, rather than just letting the information disappear down a black hole. And humility, properly misunderstood,

makes a wonderful black hole—all you have to do is admit you could be wrong. Therefore it is written: "To be humble is to take specific actions in anticipation of your own errors. To confess your fallibility and then do nothing about it is not humble; it is boasting of your modesty."

2. The Third Alternative[1]

> *"Believing in Santa Claus gives children a sense of wonder and encourages them to behave well in hope of receiving presents. If Santa-belief is destroyed by truth[2], the children will lose their sense of wonder and stop behaving nicely. Therefore, even though Santa-belief is false-to-fact, it is a Noble Lie whose net benefit should be preserved for utilitarian reasons."*

Classically, this is known as a false dilemma[3], the fallacy of the excluded middle, or the package-deal fallacy[4]. Even if we accept the underlying factual and moral premises of the above argument, it does not carry through. Even supposing that the Santa policy (encourage children to believe in Santa Claus) is better than the null policy (do nothing), it does not follow that Santa-ism is the *best of all possible alternatives*. Other policies could also supply children with a sense of wonder, such as taking them to watch a Space Shuttle launch or supplying them with science fiction novels. Likewise (if I recall correctly), offering children bribes for good behavior encourages the children to behave well *only* when adults are watching, while praise without bribes leads to unconditional good behavior.

Noble Lies are generally package-deal fallacies; and the response to a package-deal fallacy is that if we really need the supposed gain, we can construct a Third Alternative for getting it.

How can we obtain Third Alternatives? The first step in obtaining a Third Alternative is deciding to look for one, and the last step is the decision to accept it. This sounds obvious, and

1. *http://lesswrong.com/lw/hu/the_third_alternative/*
2. *http://yudkowsky.net/virtues/*
3. *http://en.wikipedia.org/wiki/False_dilemma*
4. *http://en.wikipedia.org/wiki/Package-deal_fallacy*

yet most people fail on these two steps, rather than within the search process. Where do false dilemmas come from? Some arise honestly, because superior alternatives are cognitively hard to see. But one factory for false dilemmas is justifying a questionable policy by pointing to a supposed benefit over the null action. In this case, the justifier *does not want* a Third Alternative; finding a Third Alternative would destroy the justification. The last thing a Santa-ist wants to hear is that praise works better than bribes, or that spaceships can be as inspiring as flying reindeer.

The best is the enemy of the good. If the goal is *really* to help people, then a superior alternative is cause for celebration—once we find this better strategy, we can help people more effectively. But if the goal is to justify a particular strategy *by claiming that it helps people,* a Third Alternative is an enemy argument[5], a competitor.

Modern cognitive psychology views decision-making as a search for alternatives. In real life, it's not enough to compare options, you have to generate the options in the first place. On many problems, the number of alternatives is huge, so you need a stopping criterion for the search. When you're looking to buy a house, you can't compare every house in the city; at some point you have to stop looking and decide.

But what about when our conscious motives for the search—the criteria we can admit to ourselves—don't square with subconscious influences? When we are carrying out an allegedly altruistic search, a search for an altruistic policy, and we find a strategy that benefits others but disadvantages ourselves—well, we don't stop looking *there;* we go on looking. Telling ourselves that we're looking for a strategy that brings greater altruistic benefit, of course. But suppose we find a policy that has some defensible benefit, and *also* just happens to be personally convenient? Then we stop the search at once! In fact, we'll probably *resist* any suggestion that we start looking

5. Page 148, 'Politics is the Mind-Killer'.

again—pleading lack of time, perhaps. (And yet somehow, we always have cognitive resources for coming up with justifications for our current policy.)

Beware when you find yourself arguing that a policy is *defensible* rather than *optimal;* or that it has some benefit compared to the null action, rather than the best benefit of any action.

False dilemmas are often presented to justify unethical policies that are, by some vast coincidence, very convenient. Lying, for example, is often much more convenient than telling the truth; and believing whatever you started out with is more convenient than updating. Hence the popularity of arguments for Noble Lies; it serves as a defense of a pre-existing belief—one does not find Noble Liars who calculate an optimal new Noble Lie; they keep whatever lie they started with. Better stop that search fast!

To do better[6], ask yourself straight out: *If I saw that there was a superior alternative to my current policy, would I be glad in the depths of my heart, or would I feel a tiny flash of reluctance before I let go?* If the answers are "no" and "yes", beware that you may not have searched for a Third Alternative.

Which leads into another good question to ask yourself straight out: *Did I spend five minutes with my eyes closed, brainstorming wild and creative options, trying to think of a better alternative?* It has to be five minutes by the clock, because otherwise you blink—close your eyes and open them again—and say, "Why, yes, I searched for alternatives, but there weren't any." Blinking makes a good black hole[7] down which to dump your duties. An actual, physical clock is recommended.

And those wild and creative options—were you careful not to think of a good one? Was there a secret effort from the corner of your mind to ensure that every option considered would be obviously bad?

6. *http://lesswrong.com/lw/h8/tsuyoku_naritai_i_want_to_become_stronger/*
7. Page 421, 'The Proper Use of Humility'.

It's amazing how many Noble Liars and their ilk are eager to embrace ethical violations—with all due bewailing of their agonies of conscience—when they haven't spent even five minutes by the clock looking for an alternative. There are some mental searches that we secretly wish would fail; and when the prospect of success is uncomfortable, people take the earliest possible excuse to give up.

3. Privileging the Hypothesis[1]

Suppose that the police of Largeville, a town with a million inhabitants, are investigating a murder in which there are few or no clues—the victim was stabbed to death in an alley, and there are no fingerprints and no witnesses.

Then, one of the detectives says, "Well... we have no idea who did it... no particular evidence singling out any of the million people in this city... but let's *consider the hypothesis* that this murder was committed by Mortimer Q. Snodgrass, who lives at 128 Ordinary Ln. It *could* have been him, after all."

I'll label this *the fallacy of privileging the hypothesis*. (Do let me know if it already has an official name—I can't recall seeing it described.)

Now the detective may perhaps have some form of rational evidence[2] which is not legal evidence[3] admissible in court—hearsay from an informant, for example. But if the detective does not have *some* justification *already in hand* for promoting Mortimer to the police's special attention—if the name is pulled entirely out of a hat—then Mortimer's rights are being violated.

And this is true even if the detective is not claiming that Mortimer "did" do it, but only asking the police to spend time pondering that Mortimer *might* have done it—unjustifiably promoting that particular hypothesis to attention[4]. It's human nature to look for confirmation rather than disconfirmation[5]. Suppose that three detectives each suggest their hated enemies, as names to be considered; and Mortimer is brown-haired, Frederick is black-haired, and Helen is blonde. Then a witness is found who says that the person leaving the scene was brown-

1. *http://lesswrong.com/lw/19m/privileging_the_hypothesis/*
2. *http://wiki.lesswrong.com/wiki/Rational_evidence*
3. *http://wiki.lesswrong.com/wiki/Rational_evidence*
4. *http://wiki.lesswrong.com/wiki/Locate_the_hypothesis*
5. Page 108, 'Positive Bias: Look Into the Dark'.

haired. "Aha!" say the police. "We previously had no evidence to distinguish among the possibilities, but *now* we know that Mortimer did it!"

This is related to the principle I've started calling "locating the hypothesis[6]", which is that if you have a billion boxes only one of which contains a diamond (the truth), and your detectors only provide 1 bit of evidence[7] apiece, then it takes much more evidence to promote the truth to your particular attention—to narrow it down to ten good possibilities, each deserving of our individual attention—than it does to figure out *which* of those ten possibilities is true. 27 bits to narrow it down to 10, and just another 4 bits will give us better than even odds of having the right answer. (Again, let me know if there's a more standard name for this.)

Thus the detective, in calling Mortimer to the particular attention of the police, for no reason out of a million other people, is skipping over *most of the evidence* that needs to be supplied against Mortimer.

And the detective ought to have this evidence in their possession, at the first moment when they bring Mortimer to the police's attention *at all*. It may be mere rational evidence[8] rather than legal evidence[9], but if there's *no evidence* then the detective is harassing and persecuting poor Mortimer.

During my recent diavlog with Scott Aaronson on quantum mechanics[10], I did manage to corner Scott to the extent of getting Scott to admit that there was no concrete evidence whatsoever that favors a collapse postulate[11] or single-world quantum mechanics[12]. But, said Scott, we might encounter *future* evi-

6. *http://wiki.lesswrong.com/wiki/Locate_the_hypothesis*
7. *http://wiki.lesswrong.com/wiki/Amount_of_evidence*
8. *http://wiki.lesswrong.com/wiki/Rational_evidence*
9. *http://wiki.lesswrong.com/wiki/Rational_evidence*
10. *http://www.bloggingheads.tv/diavlogs/21857*
11. *http://lesswrong.com/lw/q6/collapse_postulates/*
12. *http://lesswrong.com/lw/q8/many_worlds_one_best_guess/*

dence in favor of single-world quantum mechanics, and many-worlds still has the open question of the Born probabilities[13].

This is indeed what I would call the fallacy of privileging the hypothesis. There must be a trillion better ways to answer the Born question without adding a collapse postulate that would be the only non-linear, non-unitary, discontinous, non-differentiable, non-CPT-symmetric, non-local in the configuration space, Liouville's-Theorem-violating, privileged-space-of-simultaneity-possessing, faster-than-light-influencing, acausal, informally specified law in all of physics[14]. Something that unphysical is not worth *saying out loud* or even *thinking about as a possibility* without a rather large weight of evidence[15]—far more than the current grand total of zero.

But because of a historical accident, collapse postulates and single-world quantum mechanics are indeed on everyone's lips and in everyone's mind to be thought of, and so the open question of the Born probabilities is offered up (by Scott Aaronson no less!) as evidence that many-worlds can't yet offer a complete picture of the world. Which is taken to mean that single-world QM is still in the running somehow.

In the minds of human beings, if you can get them to think about this particular hypothesis rather than the trillion other possibilities that are no more complicated or unlikely, you really *have* done a huge chunk of the work of persuasion. Anything thought about is treated as "in the running", and if other runners seem to fall behind in the race a little, it's assumed that this runner is edging forward or even entering the lead.

And yes, this is just the same fallacy committed, on a much more blatant scale, by the theist who points out that modern science does not offer an absolutely complete explanation of the entire universe, and takes this as evidence for the existence of Jehovah. Rather than Allah, the Flying Spaghetti Monster, or

13. *http://lesswrong.com/lw/py/the_born_probabilities/*
14. *http://lesswrong.com/lw/q6/collapse_postulates/*
15. *http://wiki.lesswrong.com/wiki/Amount_of_evidence*

a trillion other gods no less complicated—never mind the space of naturalistic explanations!

To talk about "intelligent design" whenever you point to a purported flaw or open problem in evolutionary theory is, again, privileging the hypothesis—you must have evidence *already in hand* that points to intelligent design *specifically* in order to justify *raising that particular idea to our attention*, rather than a thousand others.

So that's the *sane* rule. And the corresponding anti-epistemology[16] is to talk endlessly of "possibility" and how you "can't disprove" an idea, to hope that future evidence may confirm it without presenting past evidence already in hand, to dwell and dwell on *possibilities* without evaluating possibly unfavorable evidence, to draw glowing word-pictures of confirming observations that *could* happen but haven't happened *yet*, or to try and show that piece after piece of negative evidence is "not conclusive".

Just as Occam's Razor[17] says that more complicated propositions require more evidence to believe, more complicated propositions also ought to require more work to raise to attention. Just as the principle of burdensome details[18] requires that each part of a belief be separately justified, it requires that each part be separately raised to attention.

As discussed in Perpetual Motion Beliefs[19], faith and type 2 perpetual motion machines (water—> ice cubes + electricity) have in common that they purport to *manufacture improbability from nowhere,* whether the improbability of water forming ice cubes or the improbability of arriving at correct beliefs without observation. Sometimes most of the anti-work involved in manufacturing this improbability is getting us to *pay attention* to an unwarranted belief—thinking on it, dwelling on it.

16. *http://wiki.lesswrong.com/wiki/Anti-epistemology*
17. *http://wiki.lesswrong.com/wiki/Occam%27s_razor*
18. *http://wiki.lesswrong.com/wiki/Burdensome_details*
19. *http://lesswrong.com/lw/06/perpetual_motion_beliefs/*

In large answer spaces, attention without evidence is more than halfway to belief without evidence.

Someone who spends all day thinking about whether the *Trinity* does or does not exist, rather than Allah or Thor or the Flying Spaghetti Monster, is more than halfway to Christianity. If leaving, they're less than half departed; if arriving, they're more than halfway there.

Added: An oft-encountered mode of privilege is to try to make uncertainty within a space, slop outside of that space onto the privileged hypothesis. For example, a creationist seizes on some (allegedly) debated aspect of contemporary theory, argues that scientists are *uncertain about evolution*, and then says, "We don't really know which theory is right, so maybe intelligent design is right." But the uncertainty is uncertainty *within* the realm of naturalistic theories of evolution—we have no reason to believe that we'll need to leave that realm to deal with our uncertainty, still *less* that we would jump out of the realm of standard science and land *on Jehovah in particular*. That is privileging the hypothesis—taking doubt *within* a normal space, and trying to slop doubt *out* of the normal space, onto a privileged (and usually discredited) *extremely* abnormal target.

Similarly, our uncertainty about where the Born statistics come from, should be uncertainty *within* the space of quantum theories that are continuous, linear, unitary, slower-than-light, local, causal, naturalistic, etcetera—the usual character of physical law. Some of that uncertainty might slop outside the standard space onto theories that violate *one* of these standard characteristics. It's indeed possible that we might have to think outside the box. But single-world theories violate *all* these characteristics, and there is no reason to privilege that hypothesis.

Wiki entry: Privilege the hypothesis[20].

20. *http://wiki.lesswrong.com/wiki/Privilege_the_hypothesis*

4. But There's Still A Chance, Right?[1]

Years ago, I was speaking to someone when he casually remarked that he didn't believe in evolution. And I said, "This is not the nineteenth century. When Darwin first proposed evolution, it might have been reasonable to doubt it. But this is the twenty-first century. We can *read the genes*. Humans and chimpanzees have 98% shared DNA. We *know* humans and chimps are related. It's *over*."

He said, "Maybe the DNA is just similar by coincidence."

I said, "The odds of that are something like two to the power of seven hundred and fifty million to one."

He said, "But there's still a chance, right?"

Now, there's a number of reasons my past self cannot claim a strict moral victory in this conversation. One reason is that I have no memory of whence I pulled that 2^(750,000,000) figure, though it's probably the right meta-order of magnitude. The other reason is that my past self didn't apply the concept of a calibrated confidence. Of all the times over the history of humanity that a human being has calculated odds of 2^(750,000,000):1 against something, they have undoubtedly been wrong more often than once in 2^(750,000,000) times. E.g. the shared genes estimate was revised to 95%, not 98%—and that may even apply only to the 30,000 known genes and not the entire genome, in which case it's the wrong meta-order of magnitude.

But I think the other guy's reply is still pretty funny.

I don't recall what I said in further response—probably something like "***No***"—but I remember this occasion because it brought me several insights into the laws of thought as seen by the unenlightened ones.

It first occurred to me that human intuitions were making a qualitative distinction between "No chance" and "A very tiny

1. *http://lesswrong.com/lw/ml/but_theres_still_a_chance_right/*

chance, but worth keeping track of." You can see this in the OB lottery debate[2], where someone said, "There's a big difference between zero chance of winning and epsilon chance of winning," and I replied, "No, there's an order-of-epsilon difference; if you doubt this, let epsilon equal one over googolplex."

The problem is that probability theory sometimes lets us calculate a chance which is, indeed, too tiny to be worth the mental space to keep track of it—but by that time, you've already calculated it. People mix up the map with the territory, so that on a gut level, tracking a symbolically described probability feels like "a chance worth keeping track of", even if the *referent* of the symbolic description is a number so tiny that if it was a dust speck, you couldn't see it. We can use words to describe numbers that small, but not feelings—a feeling that small doesn't exist, doesn't fire enough neurons or release enough neurotransmitters to be felt. This is why people buy lottery tickets—no one can *feel* the smallness of a probability that small.

But what I found even more fascinating was the qualitative distinction between "certain" and "uncertain" arguments, where if an argument is not certain, you're allowed to ignore it. Like, if the likelihood is zero, then you have to give up the belief, but if the likelihood is one over googol, you're allowed to keep it.

Now it's a free country and no one should put you in jail for illegal reasoning, but if you're going to ignore an argument that says the likelihood is one over googol, why not also ignore an argument that says the likelihood is zero? I mean, as long as you're ignoring the evidence anyway, why is it so much worse to ignore certain evidence than uncertain evidence?

I have often found, in life, that I have learned[3] from other people's nicely blatant bad examples, duly generalized to more subtle cases. In this case, the flip lesson is that, if you can't ig-

2. *http://lesswrong.com/lw/hm/new_improved_lottery/*
3. Page 168, 'Reversed Stupidity Is Not Intelligence'.

nore a likelihood of one over googol because you want to, you can't ignore a likelihood of 0.9 because you want to. It's all the same slippery cliff.

Consider his example if you ever you find yourself thinking, "But you can't *prove* me wrong." If you're going to ignore a probabilistic counterargument, why not ignore a proof, too?

5. The Fallacy of Gray[1]

Followup to: Tsuyoku Naritai[2], But There's Still A Chance Right?[3]

> The Sophisticate: "The world isn't black and white. No one does pure good or pure bad. It's all gray. Therefore, no one is better than anyone else."
>
> The Zetet: "Knowing only gray, you conclude that all grays are the same shade. You mock the simplicity of the two-color view, yet you replace it with a one-color view..."
>
> —Marc Stiegler, *David's Sling*

I don't know if the Sophisticate's mistake has an official name, but I call it the Fallacy of Gray. We saw it manifested in yesterday's post—the one who believed that odds of two to the power of seven hundred and fifty millon to one, against, meant "there was still a chance". All probabilities, to him, were simply "uncertain" and that meant he was licensed to ignore them if he pleased.

"The Moon is made of green cheese" and "the Sun is made of mostly hydrogen and helium" are both uncertainties, but they are not the same uncertainty.

Everything is shades of gray, but there are shades of gray so light as to be very nearly white, and shades of gray so dark as to be very nearly black. Or even if not, we can still compare shades, and say "it is darker" or "it is lighter".

Years ago, one of the strange little formative moments in my career as a rationalist was reading this paragraph from *Player of Games* by Iain M. Banks, especially the sentence in bold:

---

1. *http://lesswrong.com/lw/mm/the_fallacy_of_gray/*
2. *http://lesswrong.com/lw/h8/tsuyoku_naritai_i_want_to_become_stronger/*
3. Page 435, 'But There's Still A Chance, Right?'.

> "A guilty system recognizes no innocents. As with any power apparatus which thinks everybody's either for it or against it, we're against it. You would be too, if you thought about it. The very way you think places you amongst its enemies. This might not be your fault, because **every society imposes some of its values on those raised within it, but the point is that some societies try to maximize that effect, and some try to minimize it**. You come from one of the latter and you're being asked to explain yourself to one of the former. Prevarication will be more difficult than you might imagine; neutrality is probably impossible. You cannot choose not to have the politics you do; they are not some separate set of entities somehow detachable from the rest of your being; they are a function of your existence. I know that and they know that; you had better accept it."

Now, don't write angry comments saying that, if societies impose fewer of their values, then each succeeding generation has more work to start over from scratch. That's not what I got out of the paragraph.

What I got out of the paragraph was something which seems so obvious in retrospect that I could have conceivably picked it up in a hundred places; but something about that one paragraph made it click for me.

It was the whole notion of the Quantitative Way applied to life-problems like moral judgments and the quest for personal self-improvement. That, even if you couldn't switch something from on to off, you could still tend to increase it or decrease it.

Is this too obvious to be worth mentioning? I say it is not too obvious, for many bloggers have said of *Overcoming Bias:* "It is impossible, no one can completely eliminate bias." I don't care if the one is a professional economist, it is clear that they have not yet grokked the Quantitative Way as it applies to ev-

eryday life and matters like personal self-improvement. That which I cannot *eliminate* may be well worth *reducing*.

Or consider this exchange between Robin Hanson[4] and Tyler Cowen[5]. Robin Hanson said that he preferred to put at least 75% weight on the prescriptions of economic theory versus his intuitions: "I try to mostly just straightforwardly apply economic theory, adding little personal or cultural judgment". Tyler Cowen replied:

> In my view there is no such thing as "straightforwardly applying economic theory"... theories are always applied through our personal and cultural filters and there is no other way it can be.

Yes, but you can try to minimize that effect, or you can do things that are bound to increase it. And *if* you try to minimize it, then in many cases I don't think it's unreasonable to call the output "straightforward"—even in economics.

"Everyone is imperfect." Mohandas Gandhi was imperfect and Joseph Stalin was imperfect, but they were not the same shade of imperfection. "Everyone is imperfect" is an excellent example of replacing a two-color view with a one-color view. If you say, "No one is perfect, but *some people are less imperfect than others,*" you may not gain applause[6]; but for those who strive to do better[7], you have held out hope. No one is *perfectly* imperfect, after all.

(Whenever someone says to me, "Perfectionism is bad for you," I reply: "I think it's okay to be imperfect, but not so imperfect that other people notice.")

---

4. *http://www.overcomingbias.com/2007/12/economist-judgm.html*
5. *http://www.marginalrevolution.com/marginalrevolution/2007/12/how-my-views-di.html*
6. Page 128, 'Applause Lights'.
7. *http://lesswrong.com/lw/h8/tsuyoku_naritai_i_want_to_become_stronger/*

Likewise the folly of those who say, "Every scientific paradigm imposes some of its assumptions on how it interprets experiments," and then act like they'd proven science to occupy the same level with witchdoctoring. Every worldview imposes some of its structure on its observations, but the point is that there are worldviews which try to minimize that imposition, and worldviews which glory in it. There is no white, but there are shades of gray that are far lighter than others, and it is folly to treat them as if they were all on the same level.

If the moon has orbited the Earth these past few billion years, if you have seen it in the sky these last years, and you expect to see it in its appointed place and phase tomorrow, then that is not a certainty. And if you expect an invisible dragon[8] to heal your daughter of cancer, that too is not a certainty. But they are rather different degrees of uncertainty—this business of expecting things to happen yet again in the same way you have previously predicted to twelve decimal places, versus expecting something to happen that *violates* the order previously observed. Calling them both "faith" seems a little too un-narrow[9].

It's a most peculiar psychology—this business of "Science is based on faith too, so there!" Typically this is said by people who claim that faith is a *good* thing. Then why do they say "Science is based on faith too!" in that angry-triumphal tone, rather than as a compliment? And a rather *dangerous* compliment to give, one would think, from their perspective. If science is based on 'faith', then science is of the same kind as religion—directly comparable. If science is a religion, it is the religion that heals the sick and reveals the secrets of the stars. It would make sense to say, "The priests of science can blatantly, publicly, verifiably walk on the Moon as a faith-based miracle, and your priests' faith can't do the same." Are you sure you wish to go there, oh faithist? Perhaps, on further reflection,

8. Page 43, 'Belief in Belief'.
9. Page 58, 'The Virtue of Narrowness'.

you would prefer to retract this whole business of "Science is a religion too!"

There's a strange dynamic here: You try to purify your shade of gray, and you get it to a point where it's pretty light-toned, and someone stands up and says in a deeply offended tone, "But it's not white! It's gray!" It's one thing when someone says, "This isn't as light as you think, because of specific problems X, Y, and Z." It's a different matter when someone says angrily "It's not white! It's gray!" without pointing out any specific dark spots.

In this case, I begin to suspect psychology that is more imperfect than usual—that someone may have made a devil's bargain with their own mistakes, and now refuses to hear of any possibility of improvement. When someone finds an excuse not to try to do better, they often refuse to concede that anyone else *can* try to do better, and every mode of improvement is thereafter their enemy, and every claim that it is possible to move forward is an offense against them. And so they say in one breath proudly, "I'm glad to be gray," and in the next breath angrily, "And *you're gray too!*"

If there is no black and white, there is yet lighter and darker, and not all grays are the same.

Addendum: G[10] points us to Asimov's The Relativity of Wrong[11]: "When people thought the earth was flat, they were wrong. When people thought the earth was spherical, they were wrong. But if you think that thinking the earth is spherical is just as wrong as thinking the earth is flat, then your view is wronger than both of them put together."

10. *http://lesswrong.com/lw/mm/the_fallacy_of_gray/hho*
11. *http://chem.tufts.edu/AnswersInScience/RelativityofWrong.htm*

6. Absolute Authority[1]

Followup to: But There's Still A Chance Right?[2], The Fallacy of Gray[3]

The one comes to you and loftily says: "Science doesn't really *know* anything. All you have are *theories*—you can't know for *certain* that you're right. You scientists changed your minds about how gravity works—who's to say that tomorrow you won't change your minds about evolution?"

Behold the abyssal cultural gap[4]. If you think you can cross it in a few sentences, you are bound to be sorely disappointed.

In the world of the unenlightened ones, there is authority and un-authority. What can be trusted, can be trusted; what cannot be trusted, you may as well throw away. There are good sources of information and bad sources of information. If scientists have changed their stories ever in their history, then science cannot be a true Authority, and can never again be trusted—like a witness caught in a contradiction, or like an employee found stealing from the till.

Plus, the one takes for granted that a proponent of an idea is expected to defend it against every possible counterargument[5] and confess nothing. All claims are discounted accordingly. If even the *proponent* of science admits that science is less than perfect, why, it must be pretty much worthless.

When someone has lived their life accustomed to certainty, you can't just say to them, "Science is probabilistic, just like all other knowledge." They will accept the first half of the statement as a confession of guilt; and dismiss the second half as a flailing attempt to accuse everyone else to avoid judgment.

---

1. *http://lesswrong.com/lw/mn/absolute_authority/*
2. Page 435, 'But There's Still A Chance, Right?'.
3. Page 438, 'The Fallacy of Gray'.
4. *http://lesswrong.com/lw/kg/expecting_short_inferential_distances/*
5. Page 150, 'Policy Debates Should Not Appear One-Sided'.

You have admitted you are not trustworthy—so begone, Science, and trouble us no more!

One obvious source for this pattern of thought is religion, where the scriptures are alleged to come from God; therefore to confess any flaw in them would destroy their authority utterly; so any trace of doubt is a sin, and claiming certainty[6] is *mandatory* whether you're certain or not.

But I suspect that the traditional school regimen also has something to do with it. The teacher tells you certain things, and you have to believe them, and you have to recite them back on the test. But when a student makes a suggestion in class, you don't have to go along with it—you're free to agree or disagree (it seems) and no one will punish you.

This experience, I fear, maps the domain of belief onto the social domains of *authority,* of *command*, of *law*. In the social domain, there is a qualitative difference between absolute laws and nonabsolute laws, between commands and suggestions, between authorities and unauthorities. There seems to be strict knowledge and unstrict knowledge, like a strict regulation and an unstrict regulation. Strict authorities must be yielded to, while unstrict suggestions can be obeyed or discarded as a matter of personal preference. And Science, since it confesses itself to have a possibility of error, must belong in the second class.

(I note in passing that I see a certain similarity to they who think that if you don't get an Authoritative probability written on a piece of paper from the teacher in class, or handed down from some similar Unarguable Source, then your uncertainty[7] is not a matter for Bayesian probability theory. Someone might—*gasp!*—argue with your estimate of the prior probability. It thus seems to the not-fully-enlightened ones that Bayesian priors belong to the class of beliefs proposed by students, and not the class of beliefs commanded you by teachers—it is not proper *knowledge*.)

6. Page 50, 'Professing and Cheering'.
7. Page 55, 'Focus Your Uncertainty'.

The abyssal cultural gap between the Authoritative Way and the Quantitative Way is rather annoying to those of us staring across it from the rationalist side. Here is someone who believes they have knowledge *more* reliable than science's mere probabilistic guesses—such as the guess that the moon will rise in its appointed place and phase tomorrow, just like it has every observed night since the invention of astronomical record-keeping, and just as predicted by physical theories whose previous predictions have been successfully confirmed to fourteen decimal places. And what is this knowledge that the unenlightened ones set above ours, and why? It's probably some musty old scroll that has been contradicted eleventeen ways from Sunday, and from Monday, and from every day of the week. Yet this is more reliable than Science (they say) because it never admits to error, never changes its mind, no matter how often it is contradicted. They toss around the word "certainty" like a tennis ball, using it as lightly as a feather—while scientists are weighed down by dutiful doubt, struggling to achieve even a modicum of probability. "I'm perfect," they say without a care in the world, "I must be so far above *you,* who must still struggle to improve yourselves."

There is nothing simple[8] you can say to them—no *fast* crushing rebuttal. By thinking carefully, you may be able to win over the audience, if this is a public debate. Unfortunately you cannot just blurt out[9], "Foolish mortal, the Quantitative Way is beyond your comprehension, and the beliefs you lightly name 'certain' are less assured than the least of our mighty hypotheses." It's a difference of *life-gestalt* that isn't easy to describe in words at all, let alone quickly.

What might you try, rhetorically, in front of an audience? Hard to say... maybe:

- "The power of science comes from having the ability to change our minds and admit we're wrong. If you've

8. *http://lesswrong.com/lw/kg/expecting_short_inferential_distances/*
9. *http://lesswrong.com/lw/kg/expecting_short_inferential_distances/*

never admitted you're wrong, it doesn't mean you've made fewer mistakes."

- "Anyone can *say* they're absolutely certain. It's a bit harder to never, ever make any mistakes. Scientists understand the difference, so they don't say they're absolutely certain. That's all. It doesn't mean that they have any specific reason to doubt a theory—absolutely every scrap of evidence can be going the same way, all the stars and planets lined up like dominos in support of a single hypothesis, and the scientists still won't say they're absolutely sure, because they've just got higher standards. It doesn't mean scientists are less *entitled* to certainty than, say, the politicians who always seem so sure of everything."
- "Scientists don't use the phrase 'not absolutely certain' the way you're used to from regular conversation. I mean, suppose you went to the doctor, and got a blood test, and the doctor came back and said, 'We ran some tests, and it's not absolutely certain that you're not made out of cheese, and there's a non-zero chance that twenty fairies made out of sentient chocolate are singing the 'I love you' song from Barney inside your lower intestine.' Run for the hills, your doctor needs a doctor. When a scientist says the same thing, it means that he thinks the probability is so tiny that you couldn't see it with an electron microscope, but he's willing to see the evidence in the extremely unlikely event that you have it."
- "Would you be willing to change your mind about the things you call 'certain' if you saw enough evidence? I mean, suppose that God himself descended from the clouds and told you that your whole religion was true except for the Virgin Birth. If that would change your mind, you can't say you're absolutely certain of the Virgin Birth. For technical reasons of probability theory, if it's theoretically possible for you to change

> your mind about something, it can't have a probability exactly equal to one. The uncertainty might be smaller than a dust speck, but it has to be there. And if you wouldn't change your mind even if God told you otherwise, then you have a problem with refusing to admit you're wrong that transcends anything a mortal like me can say to you, I guess."

But, in a way, the more interesting question is what you say to someone *not* in front of an audience. How do you begin the long process of teaching someone to live in a universe without certainty?

I think the first, beginning step should be understanding that you *can* live without certainty—that *if, hypothetically speaking,* you couldn't be certain of anything, it would not deprive you of the ability to make moral or factual distinctions. To paraphrase Lois Bujold, "Don't push harder, lower the resistance."

One of the common *defenses* of Absolute Authority is something I call "The Argument From The Argument From Gray", which runs like this:

- *Moral relativists say:*
  - The world isn't black and white, therefore:
  - Everything is gray, therefore:
  - No one is better than anyone else, therefore:
  - I can do whatever I want and you can't stop me bwahahaha.
- But we've got to be able to stop people from committing murder.
- Therefore there has to be some way of being absolutely certain, or the moral relativists win.

Reversed stupidity[10] is not intelligence. You can't arrive at a correct answer by reversing *every single* line of an argument that ends with a bad conclusion—it gives the fool too much detailed control over you. Every single line[11] must be correct for

10. Page 168, 'Reversed Stupidity Is Not Intelligence'.

a mathematical argument to carry. And it doesn't follow, from the fact that moral relativists say "The world isn't black and white", that this is false, any more than it follows from Stalin's belief that 2 + 2 = 4 that "2 + 2 = 4" is false. The error (and it only takes one) is in the leap from the two-color view to the single-color view, that all grays are the same shade.

It would concede far too much (indeed, concede the whole argument) to agree with the premise that you need absolute knowledge of absolutely good options and absolutely evil options in order to be moral. You can have uncertain knowledge of relatively better and relatively worse options, and still choose. It should be routine, in fact, not something to get all dramatic about.

I mean, yes, if you have to choose between two alternatives A and B, and you somehow succeed in establishing knowably certain well-calibrated 100% confidence that A is absolutely and entirely desirable and that B is the sum of everything evil and disgusting, then this is a *sufficient* condition for choosing A over B. It is not a *necessary* condition.

Oh, and: Logical fallacy: Appeal to consequences of belief.[12]

Let's see, what else do they need to know? Well, there's the entire rationalist culture which says that doubt, questioning, and confession of error are not terrible shameful things.

There's the whole notion of gaining information by *looking at things,* rather than being proselytized. When you look at things harder, sometimes you find out that they're different from what you thought they were at first glance; but it doesn't mean that Nature lied to you, or that you should give up on seeing.

Then there's the concept of a calibrated confidence—that "probability" isn't the same concept as the little progress bar in your head that measures your emotional commitment to an

11. *http://lesswrong.com/lw/jk/burdensome_details/*
12. *http://www.nizkor.org/features/fallacies/appeal-to-consequences.html*

idea. It's more like a measure of how often, pragmatically, in real life, people in a certain state of belief say things that are actually true. If you take one hundred people and ask them to list one hundred statements of which they are "absolutely certain", how many will be correct? Not one hundred.

If anything, the statements that people are really fanatic about are *far less* likely to be correct than statements like "the Sun is larger than the Moon" that seem too obvious to get excited about. For every statement you can find of which someone is "absolutely certain", you can probably find someone "absolutely certain" of its opposite, because such fanatic professions of belief do not arise in the absence of opposition. So the little progress bar in people's heads that measures their emotional commitment to a belief does not translate well into a calibrated confidence—it doesn't even behave monotonically.

As for "absolute certainty"—well, if you say that something is 99.9999% probable, it means you think you could make *one million* equally strong independent statements, *one after the other,* over the course of a solid year or so, and be wrong, on average, around once. This is incredible enough. (It's amazing to realize we can actually *get* that level of confidence for "Thou shalt not win the lottery.[13]") So let us say nothing of probability 1.0. Once you realize you don't *need* probabilities of 1.0 to get along in life, you'll realize how absolutely ridiculous it is to think you could ever get to 1.0 with a human brain. A probability of 1.0 isn't just certainty, it's *infinite certainty.*

In fact, it seems to me that to prevent public misunderstanding, maybe scientists should go around saying "We are not INFINITELY certain" rather than "We are not certain". For the latter case, in ordinary discourse, suggests you know some specific reason for doubt.

13. *http://lesswrong.com/lw/hm/new_improved_lottery/*

7. Infinite Certainty[1]

Followup to: How To Convince Me That 2 + 2 = 3[2], Absolute Authority[3]

In Absolute Authority[4], I argued that you don't *need* infinite certainty: "If you have to choose between two alternatives A and B, and you somehow succeed in establishing knowably certain well-calibrated 100% confidence that A is absolutely and entirely desirable and that B is the sum of everything evil and disgusting, then this is a *sufficient* condition for choosing A over B. It is not a *necessary* condition... You can have uncertain knowledge of relatively better and relatively worse options, and still choose. It should be routine, in fact."

However, might there not be *some* propositions in which we are entitled to infinite confidence? What about the proposition that 2 + 2 = 4[5]?

We must distinguish between the the map and the territory[6]. Given the seeming absolute stability and universality of physical laws[7], it's possible that never, in the whole history of the universe, has any particle exceeded the local lightspeed limit. That is, the lightspeed limit may be, not just true 99% of the time, or 99.9999% of the time, or (1—1/googolplex) of the time, but simply *always and absolutely true*.

But whether we can ever have *absolute confidence* in the lightspeed limit is a whole 'nother question. The map is not the territory.

It may be entirely and wholly true that a student plagiarized their assignment[8], but whether you have any knowledge of this

1. *http://lesswrong.com/lw/mo/infinite_certainty/*
2. Page 26, 'How to Convince Me That 2 + 2 = 3'.
3. Page 443, 'Absolute Authority'.
4. Page 443, 'Absolute Authority'.
5. Page 26, 'How to Convince Me That 2 + 2 = 3'.
6. *http://yudkowsky.net/bayes/truth.html*
7. *http://lesswrong.com/lw/hr/universal_law/*

fact at all—let alone *absolute* confidence in the belief—is a separate issue. If you flip a coin and then don't look at it, it may be completely true that the coin is showing heads, and you may be completely unsure of whether the coin is showing heads or tails. A degree of uncertainty is not the same as a degree of truth or a frequency of occurrence.

The same holds for mathematical truths. It's questionable whether the statement "2 + 2 = 4" or "In Peano arithmetic, SS0 + SS0 = SSSS0" can be said to be *true* in any purely abstract sense, apart from physical systems that seem to behave in ways similar to the Peano axioms. Having said this, I will charge right ahead and guess that, in whatever sense "2 + 2 = 4" is true at all, it is always and precisely true, not just roughly true ("2 + 2 actually equals 4.0000004") or true 999,999,999,999 times out of 1,000,000,000,000.

I'm not totally sure what "true" should mean in this case, but I stand by my guess. The credibility of "2 + 2 = 4 is always true" far exceeds the credibility of any particular philosophical position on what "true", "always", or "is" means in the statement above.

This doesn't mean, though, that I have *absolute confidence* that 2 + 2 = 4. See the previous discussion on how to convince me that $2 + 2 = 3^9$, which could be done using much the same sort of evidence that convinced me that 2 + 2 = 4 in the first place. I could have hallucinated all that previous evidence, or I could be misremembering it. In the annals of neurology there are stranger brain dysfunctions than this.

So if we attach some probability to the statement "2 + 2 = 4", then what should the probability be? What you seek to attain in a case like this is good calibration—statements to which you assign "99% probability" come true 99 times out of 100. This is actually a hell of a lot more difficult than you might think. Take a hundred people, and ask each of them to make ten statements

8. *http://ansuz.sooke.bc.ca/bonobo-conspiracy/?i=369*
9. Page 26, 'How to Convince Me That 2 + 2 = 3'.

of which they are "99% confident". Of the 1000 statements, do you think that around 10 will be wrong?

I am not going to discuss the actual experiments that have been done on calibration—you can find them in my book chapter[10]—because I've seen that when I blurt this out to people without proper preparation, they thereafter use it as a Fully General Counterargument[11], which somehow leaps to mind whenever they have to discount the confidence of someone whose opinion they dislike, and fails to be available when they consider their own opinions. So I try not to talk about the experiments on calibration except as part of a structured presentation of rationality that includes warnings against motivated skepticism.

But the observed calibration of human beings who say they are "99% confident" is not 99% accuracy.

Suppose you say that you're 99.99% confident that 2 + 2 = 4. Then you have just asserted that you could make 10,000 *independent* statements, in which you repose equal confidence, and be wrong, on average, around once. Maybe for 2 + 2 = 4 this extraordinary degree of confidence would be possible: "2 + 2 = 4" extremely simple, and mathematical as well as empirical, and widely believed socially (not with passionate affirmation but just quietly taken for granted). So maybe you really could get up to 99.99% confidence on this one.

I don't think you could get up to 99.99% confidence for assertions like "53 is a prime number". Yes, it seems likely, but by the time you tried to set up protocols that would let you assert 10,000 *independent* statements of this sort—that is, not just a set of statements about prime numbers, but a new protocol each time—you would fail more than once. Peter de Blanc has an amusing anecdote on this point, which he is welcome to retell in the comments.

10. *http://intelligence.org/Biases.pdf*
11. Page 333, 'Knowing About Biases Can Hurt People'.

Yet the map is not the territory: if I say that I am 99% confident that 2 + 2 = 4, it doesn't mean that I think "2 + 2 = 4" is true to within 99% precision, or that "2 + 2 = 4" is true 99 times out of 100. The proposition in which I repose my confidence is the proposition that "2 + 2 = 4 is always and exactly true", not the proposition "2 + 2 = 4 is mostly and usually true".

As for the notion that you could get up to 100% confidence in a mathematical proposition—well, really now! If you say 99.9999% confidence, you're implying that you could make *one million* equally fraught statements, one after the other, and be wrong, on average, about once. That's around a solid year's worth of talking, if you can make one assertion every 20 seconds and you talk for 16 hours a day.

Assert 99.9999999999% confidence, and you're taking it up to a trillion. Now you're going to talk for a hundred human lifetimes, and not be wrong even once?

Assert a confidence of (1—1/googolplex) and your ego far exceeds that of mental patients who think they're God.

And a googolplex is a lot smaller than even relatively small inconceivably huge numbers like $3\text{^^^}3$[12].

But even a confidence of (1 - 1/3^^^3) isn't all that much closer to **PROBABILITY 1** than being 90% sure of something.

If all else fails, the hypothetical Dark Lords of the Matrix, who are *right now* tampering with your brain's credibility assessment of *this very sentence,* will bar the path and defend us from the scourge of infinite certainty.

Am I absolutely sure of that?

Why, of course not.

As Rafal Smigrodski once said:

> "I would say you should be able to assign a less than 1 certainty level to the mathematical concepts which

12. *http://lesswrong.com/lw/kn/torture_vs_dust_specks/*

are necessary to derive Bayes' rule itself, and still practically use it. I am not totally sure I have to be always unsure. Maybe I could be legitimately sure about something. But once I assign a probability of 1 to a proposition, I can never undo it. No matter what I see or learn, I have to reject everything that disagrees with the axiom. I don't like the idea of not being able to change my mind, ever."

8. 0 And 1 Are Not Probabilities[1]

Followup to: Infinite Certainty[2] [3]

1, 2, and 3 are all integers, and so is -4. If you keep counting up, or keep counting down, you're bound to encounter a whole lot more integers. You will not, however, encounter anything called "positive infinity" or "negative infinity", so these are not integers.

Positive and negative infinity are not integers, but rather special symbols for talking about the behavior of integers. People sometimes say something like, "5 + infinity = infinity", because if you start at 5 and keep counting up without ever stopping, you'll get higher and higher numbers without limit. But it doesn't follow from this that "infinity - infinity = 5". You can't count up from 0 without ever stopping, and then count down without ever stopping, and then find yourself at 5 when you're done.

From this we can see that infinity is not only not-an-integer, it doesn't even *behave* like an integer. If you unwisely try to mix up infinities with integers, you'll need all sorts of special new inconsistent-seeming behaviors which you don't need for 1, 2, 3 and other *actual* integers.

Even though infinity isn't an integer, you don't have to worry about being left at a loss for numbers. Although people have seen five sheep, millions of grains of sand, and septillions of atoms, no one has ever counted an infinity of anything. The same with continuous quantities—people have measured dust specks a millimeter across, animals a meter across, cities kilometers across, and galaxies thousands of lightyears across, but no one has ever measured anything an infinity across. In the real world, you don't *need* a whole lot of infinity.

---

1. *http://lesswrong.com/lw/mp/0_and_1_are_not_probabilities/*
2. Page 450, 'Infinite Certainty'.
3. Page 443, 'Absolute Authority'.

(I should note for the more sophisticated readers in the audience that they do not need to write me with elaborate explanations of, say, the difference between ordinal numbers and cardinal numbers. Yes, I possess various advanced set-theoretic definitions of infinity, but I don't see a good use for them in probability theory. See below.)

In the usual way of writing probabilities, probabilities are between 0 and 1. A coin might have a probability of 0.5 of coming up tails, or the weatherman might assign probability 0.9 to rain tomorrow.

This isn't the only way of writing probabilities, though. For example, you can transform probabilities into odds via the transformation O = (P / (1 - P)). So a probability of 50% would go to odds of 0.5/0.5 or 1, usually written 1:1, while a probability of 0.9 would go to odds of 0.9/0.1 or 9, usually written 9:1. To take odds back to probabilities you use P = (O / (1 + O)), and this is perfectly reversible, so the transformation is an isomorphism—a two-way reversible mapping. Thus, probabilities and odds are isomorphic, and you can use one or the other according to convenience.

For example, it's more convenient to use odds when you're doing Bayesian updates. Let's say that I roll a six-sided die: If any face except 1 comes up, there's an 10% chance of hearing a bell, but if the face 1 comes up, there's a 20% chance of hearing the bell. Now I roll the die, and hear a bell. What are the *odds* that the face showing is 1? Well, the prior odds are 1:5 (corresponding to the real number 1/5 = 0.20) and the likelihood ratio is 0.2:0.1 (corresponding to the real number 2) and I can just multiply these two together to get the posterior odds 2:5 (corresponding to the real number 2/5 or 0.40). Then I convert back into a probability, if I like, and get (0.4 / 1.4) = 2/7 = ~29%.

So odds are more manageable for Bayesian updates—if you use probabilities, you've got to deploy Bayes's Theorem[4] in its

4. *http://yudkowsky.net/rational/bayes*

complicated version. But probabilities are more convenient for answering questions like "If I roll a six-sided die, what's the chance of seeing a number from 1 to 4?" You can add up the probabilities of 1/6 for each side and get 4/6, but you can't add up the odds ratios of 0.2 for each side and get an odds ratio of 0.8.

Why am I saying all this? To show that "odd ratios" are just as legitimate a way of mapping uncertainties onto real numbers as "probabilities". Odds ratios are more convenient for some operations, probabilities are more convenient for others. A famous proof called Cox's Theorem (plus various extensions and refinements thereof) shows that all ways of representing uncertainties that obey some reasonable-sounding constraints, end up isomorphic to each other.

Why does it matter that odds ratios are just as legitimate as probabilities? Probabilities as ordinarily written are between 0 and 1, and both 0 and 1 look like they ought to be readily reachable quantities—it's easy to see 1 zebra or 0 unicorns. But when you transform probabilities onto odds ratios, 0 goes to 0, but 1 goes to positive infinity. Now absolute truth doesn't look like it should be so easy to reach.

A representation that makes it even simpler to do Bayesian updates is the log odds—this is how E. T. Jaynes recommended thinking about probabilities. For example, let's say that the prior probability of a proposition is 0.0001—this corresponds to a log odds of around -40 decibels. Then you see evidence that seems 100 times more likely if the proposition is true than if it is false. This is 20 decibels of evidence. So the posterior odds are around -40 db + 20 db = -20 db, that is, the posterior probability is ~0.01.

When you transform probabilities to log odds, 0 goes onto negative infinity and 1 goes onto positive infinity. Now both infinite certainty and infinite improbability seem a bit more out-of-reach.

In probabilities, 0.9999 and 0.99999 seem to be only 0.00009 apart, so that 0.502 is much further away from 0.503 than 0.9999 is from 0.99999. To get to probability 1 from probability 0.99999, it seems like you should need to travel a distance of merely 0.00001.

But when you transform to odds ratios, 0.502 and .503 go to 1.008 and 1.012, and 0.9999 and 0.99999 go to 9,999 and 99,999. And when you transform to log odds, 0.502 and 0.503 go to 0.03 decibels and 0.05 decibels, but 0.9999 and 0.99999 go to 40 decibels and 50 decibels.

When you work in log odds, **the distance between any two degrees of uncertainty equals the amount of evidence you would need to go from one to the other**. That is, the log odds gives us a natural measure of spacing among degrees of confidence.

Using the log odds exposes the fact that reaching infinite certainty[5] requires infinitely strong evidence, just as infinite absurdity requires infinitely strong counterevidence.

Furthermore, all sorts of standard theorems in probability have special cases if you try to plug 1s or 0s into them—like what happens if you try to do a Bayesian update on an observation to which you assigned probability 0.

So I propose that it makes sense to say that 1 and 0 are not in the probabilities; just as negative and positive infinity, which do not obey the field axioms, are not in the real numbers.

The main reason this would upset probability theorists is that we would need to rederive theorems previously obtained by assuming that we can marginalize over a joint probability by adding up all the pieces and having them sum to 1.

However, in the real world, when you roll a die, it doesn't literally have infinite certainty[6] of coming up some number between 1 and 6. The die might land on its edge; or get struck by

5. Page 450, 'Infinite Certainty'.
6. Page 450, 'Infinite Certainty'.

a meteor; or the Dark Lords of the Matrix might reach in and write "37" on one side.

If you made a magical symbol to stand for "all possibilities I haven't considered", then you could marginalize over the events including this magical symbol, and arrive at a magical symbol "T" that stands for infinite certainty.

But I would rather ask whether there's some way to derive a theorem without using magic symbols with special behaviors. That would be more elegant. Just as there are mathematicians who refuse to believe in double negation or infinite sets, I would like to be a probability theorist who doesn't believe in absolute certainty.

PS: Here's Peter de Blanc's "mathematical certainty" anecdote[7]. (I told him not to do it again.)

7. *http://www.spaceandgames.com/?p=27*

Letting Go

1. Feeling Rational[1]

A popular belief about "rationality" is that rationality opposes all emotion—that all our sadness and all our joy are automatically anti-logical by virtue of being *feelings*. Yet strangely enough, I can't find any theorem of probability theory which proves that I should appear ice-cold and expressionless.

So is rationality orthogonal to feeling? No; our emotions arise from our models of reality. If I believe that my dead brother has been discovered alive, I will be happy; if I wake up and realize it was a dream, I will be sad. P. C. Hodgell said: "That which can be destroyed by the truth should be." My dreaming self's happiness was opposed by truth. My sadness on waking is rational; there is no truth which destroys it.

Rationality begins by asking how-the-world-is, but spreads virally to any other thought which depends on how we think the world is. By talking about your beliefs about "how-the-world-is", I mean anything you believe is out there in reality, anything that either does or does not exist, any member of the class "things that can make other things happen". If you believe that there is a goblin in your closet that ties your shoe's laces together, then this is a belief about how-the-world-is. Your shoes are real—you can pick them up. If there's something out there which can reach out and tie your shoelaces together, it must be real too, part of the vast web of causes and effects we call the "universe".

Feeling angry at the goblin who tied your shoelaces involves a state of mind that is not *just* about how-the-world-is. Suppose that, as a Buddhist or a lobotomy patient or just a very phlegmatic person, finding your shoelaces tied together didn't make you angry. This wouldn't affect what you expected to see in the world—you'd still expect to open up your closet and find your shoelaces tied together. Your anger or calm shouldn't affect your best guess here, because what happens in your closet

1. *http://lesswrong.com/lw/hp/feeling_rational/*

does not depend on your emotional state of mind; though it may take some effort to think that clearly.

But the angry feeling is tangled up with a state of mind that *is* about how-the-world-is; you become angry *because* you think the goblin tied your shoelaces. The criterion of rationality spreads virally, from the initial question of whether or not a goblin tied your shoelaces, to the resulting anger.

Becoming more rational—arriving at better estimates of how-the-world-is—can diminish feelings *or intensify* them. Sometimes we run away from strong feelings by denying the facts, by flinching away from the view of the world that gave rise to the powerful emotion. If so, then as you study the skills of rationality and train yourself not to deny facts, your feelings will become stronger.

In my early days I was never quite certain whether it was *all right* to feel things strongly—whether it was allowed, whether it was proper. I do not think this confusion arose only from my youthful misunderstanding of rationality. I have observed similar troubles in people who do not even aspire to be rationalists; when they are happy, they wonder if they are really allowed to be happy, and when they are sad, they are never quite sure whether to run away from the emotion or not. Since the days of Socrates at least, and probably long before, the way to appear cultured and sophisticated has been to never let anyone see you care strongly about anything. It's *embarrassing* to feel—it's just not done in polite society. You should see the strange looks I get when people realize how much I care about rationality. It's not the unusual subject, I think, but that they're not used to seeing sane adults who visibly care about *anything*.

But I know, now, that there's nothing wrong with feeling strongly. Ever since I adopted the rule of "That which can be destroyed by the truth should be," I've also come to realize "That which the truth nourishes should thrive." When something good happens, I am happy, and there is no confusion in my mind about whether it is rational for me to be happy.

When something terrible happens[2], I do not flee my sadness by searching for fake consolations and false silver linings. I visualize the past and future of humankind, the tens of billions of deaths over our history, the misery and fear, the search for answers, the trembling hands reaching upward out of so much blood, what we could become someday when we make the stars our cities, all that darkness and all that light—I know that I can never truly understand it, and I haven't the words to say. Despite all my philosophy I am still embarrassed to confess strong emotions, and you're probably uncomfortable hearing them. But I know, now, that it is rational to feel.

2. *http://yudkowsky.net/yehuda.html*

2. The Importance of Saying "Oops"[1]

I just finished reading a history of Enron's downfall, *The Smartest Guys in the Room,* which hereby wins my award for "Least Appropriate Book Title".

An unsurprising feature of Enron's slow rot and abrupt collapse was that the executive players never admitted to having made a *large* mistake. When catastrophe #247 grew to such an extent that it required an actual policy change, they would say "Too bad that didn't work out—it was such a good idea—how are we going to hide the problem on our balance sheet?" As opposed to, "It now seems obvious in retrospect that it was a mistake from the beginning." As opposed to, "I've been stupid." There was never a watershed moment, a moment of humbling realization, of acknowledging a *fundamental* problem. After the bankruptcy, Jeff Skilling, the former COO and brief CEO of Enron, declined his own lawyers' advice to take the Fifth Amendment; he testified before Congress that Enron had been a *great* company.

Not every change is an improvement, but every improvement is necessarily a change. If we only admit small local errors, we will only make small local changes. The motivation for a *big* change comes from acknowledging a *big* mistake.

As a child I was raised on equal parts science and science fiction, and from Heinlein to Feynman I learned the tropes of Traditional Rationality: Theories must be bold and expose themselves to falsification; be willing to commit the heroic sacrifice of giving up your own ideas when confronted with contrary evidence; play nice in your arguments; try not to deceive yourself; and other fuzzy verbalisms.

A traditional rationalist upbringing tries to produce arguers who will concede to contrary evidence *eventually*— there should be *some* mountain of evidence sufficient to move you. This is not trivial; it distinguishes science from religion. But

1. *http://lesswrong.com/lw/i9/the_importance_of_saying_oops/*

there is less focus on *speed,* on giving up the fight *as quickly as possible,* integrating evidence *efficiently* so that it only takes a *minimum* of contrary evidence to destroy your cherished belief.

I was raised in Traditional Rationality, and thought myself quite the rationalist. I switched to Bayescraft (Laplace/Jaynes/Tversky/Kahneman) in the aftermath of... well, it's a long story. Roughly, I switched because I realized that Traditional Rationality's fuzzy verbal tropes had been insufficient to prevent me from making a large mistake.

After I had finally and fully admitted my mistake, I looked back upon the path that had led me to my Awful Realization. And I saw that I had made a series of small concessions, minimal concessions, grudgingly conceding each millimeter of ground, realizing as little as possible of my mistake on each occasion, admitting failure only in small tolerable nibbles. I could have moved so much faster, I realized, if I had simply screamed *"OOPS!"*

And I thought: *I must raise the level of my game.*

There is a *powerful advantage* to admitting you have made a *large* mistake. It's painful. It can also change your whole life.

It is *important* to have the watershed moment, the moment of humbling realization. To acknowledge a *fundamental* problem, not divide it into palatable bite-size mistakes.

Do not indulge in drama and become proud of admitting errors[2]. It is surely superior to get it right the first time. But if you do make an error, better by far to see it all at once. Even hedonically, it is better to take one large loss than many small ones. The alternative is stretching out the battle with yourself over years. The alternative is Enron.

Since then I have watched others making their own series of minimal concessions, grudgingly conceding each millimeter of ground; never confessing a global mistake where a local one will do; always learning as little as possible from each error. What they could fix in one fell swoop voluntarily, they transform into

2. *http://lesswrong.com/lw/h8/tsuyoku_naritai_i_want_to_become_stronger/*

tiny local patches they must be argued into. Never do they say, after confessing one mistake, *I've been a fool.* They do their best to minimize their embarrassment by saying *I was right in principle*, or *It could have worked,* or *I still want to embrace the true essence of whatever-I'm-attached-to.* Defending their pride in this passing moment, they ensure they will again make the same mistake, and again need to defend their pride.

Better to swallow the entire bitter pill in one terrible gulp.

3. The Crackpot Offer[1]

When I was very young—I think thirteen or maybe fourteen—I thought I had found a disproof of Cantor's Diagonal Argument, a famous theorem which demonstrates that the real numbers outnumber the rational numbers. Ah, the dreams of fame and glory that danced in my head!

My idea was that since each whole number can be decomposed into a bag of powers of 2, it was possible to map the whole numbers onto the set of subsets of whole numbers simply by writing out the binary expansion. 13, for example, 1101, would map onto {0, 2, 3}. It took a whole week before it occurred to me that perhaps I should *apply* Cantor's Diagonal Argument to my clever construction, and of course it found a counterexample—the binary number ...1111, which does not correspond to any finite whole number.

So I found this counterexample, and saw that my attempted disproof was false, along with my dreams of fame and glory.

I was initially a bit disappointed.

The thought went through my mind: "I'll get that theorem eventually! *Someday* I'll disprove Cantor's Diagonal Argument, even though my first try failed!" I resented the theorem for being obstinately true, for depriving me of my fame and fortune, and I began to look for other disproofs.

And then I realized something. I realized that I had made a mistake, and that, now that I'd spotted my mistake, there was absolutely no reason to suspect the strength of Cantor's Diagonal Argument any more than other major theorems of mathematics.

I saw then very clearly that I was being offered the opportunity to become a math crank, and to spend the rest of my life writing angry letters in green ink to math professors. (I'd read a book once about math cranks.)

1. *http://lesswrong.com/lw/j8/the_crackpot_offer/*

I did not wish this to be my future, so I gave a small laugh, and let it go. I waved Cantor's Diagonal Argument on with all good wishes, and I did not question it again.

And I don't remember, now, if I thought this at the time, or if I thought it afterward... but what a terribly unfair test to visit upon a child of thirteen. That I had to be that rational, already, at that age, or fail.

The smarter you are, the younger you may be, the first time you have what looks to you like a really revolutionary idea. I was lucky in that I saw the mistake myself; that it did not take another mathematician to point it out to me, and perhaps give me an outside source to blame. I was lucky in that the disproof was simple enough for me to understand. Maybe I would have recovered eventually, otherwise. I've recovered from much worse, as an adult. But if I had gone wrong that early, would I ever have developed that skill?

I wonder how many people writing angry letters in green ink were thirteen when they made that first fatal misstep. I wonder how many were promising minds before then.

I made a mistake. That was all. I was not *really right, deep down;* I did not win a moral victory; I was not displaying ambition or skepticism or any other wondrous virtue; it was not a reasonable error; I was not half right or even the tiniest fraction right. I thought a thought I would never have thought if I had been wiser, and that was all there ever was to it.

If I had been unable to admit this to myself, if I had reinterpreted my mistake as virtuous, if I had insisted on being at least a *little* right for the sake of pride, then I would not have let go. I would have gone on looking for a flaw in the Diagonal Argument. And, sooner or later, I might have found one.

Until you admit you were wrong[2], you cannot get on with your life; your self-image will still be bound to the old mistake.

Whenever you are tempted to hold on to a thought you would never have thought if you had been wiser, you are being

2. Page 466, 'The Importance of Saying "Oops"'.

offered the opportunity to become a crackpot—even if you never write any angry letters in green ink. If no one bothers to argue with you, or if you never tell anyone your idea, you may still be a crackpot. It's the *clinging* that defines it.

It's not true. It's not true deep down. It's not half-true or even a little true. It's nothing but a thought you should never have thought. Not every cloud has a silver lining. Human beings make mistakes, and not all of them are disguised successes. Human beings make mistakes; it happens, that's all. Say "oops[3]", and get on with your life.

3. Page 466, 'The Importance of Saying "Oops"'.

4. Just Lose Hope Already[1]

Casey Serin, a 24-year-old web programmer with no prior experience in real estate, owes banks 2.2 million dollars[2] after lying on mortgage applications in order to simultaneously buy 8 different houses in different states. He took cash out of the mortgage (applied for larger amounts than the price of the house) and spent the money on living expenses and real-estate seminars. He was expecting the market to go up, it seems.

That's not even the sad part. The sad part is that *he still hasn't given up.* Casey Serin does not accept defeat. He refuses to declare bankruptcy, or get a job; he still thinks[3] he can make it big in real estate. He went on spending money on seminars. He tried to take out a mortgage on a 9th house. He hasn't *failed,* you see, he's just had a *learning experience.*

That's what happens when you refuse to lose hope.

While this behavior may seem to be merely stupid, it also puts me in mind of two Nobel-Prize-winning economists...

...namely Merton and Scholes of Long-Term Capital Management[4].

While LTCM raked in giant profits over its first three years, in 1998 the inefficiences that LTCM were exploiting had started to vanish—other people knew about the trick, so it stopped working.

LTCM refused to lose hope. Addicted to 40% annual returns, they borrowed more and more leverage to exploit tinier and tinier margins. When everything started to go wrong for LTCM, they had equity of $4.72 billion, leverage of $124.5 billion, and derivative positions of $1.25 trillion.

---

1. *http://lesswrong.com/lw/gx/just_lose_hope_already/*
2. *http://iamfacingforeclosure.com/*
3. *http://iamfacingforeclosure.com/147/top-5-advice-they-say-i-have-been-ignoring-2/*
4. *http://en.wikipedia.org/wiki/Long-Term_Capital_Management*

Every profession has a different way to be smart—different skills to learn and rules to follow. You might therefore think that the study of "rationality", as a general discipline, wouldn't have much to contribute to real-life success. And yet it seems to me that *how to not be stupid* has a great deal in common across professions. If you set out to teach someone *how to not turn little mistakes into big mistakes*, it's nearly the same art whether in hedge funds or romance, and one of the keys is this: Be ready to admit you lost.

5. The Proper Use of Doubt[1]

Once, when I was holding forth upon the Way[2], I remarked upon how most organized belief systems exist to *flee from doubt*. A listener replied to me that the Jesuits must be immune from this criticism, because they practice organized doubt: their novices, he said, are told to doubt Christianity; doubt the existence of God; doubt if their calling is real; doubt that they are suitable for perpetual vows of chastity and poverty. And I said: *Ah, but they're supposed to overcome these doubts, right?* He said: *No, they are to doubt that perhaps their doubts may grow and become stronger.*

Googling failed to confirm or refute these allegations. (If anyone in the audience can help, I'd be much obliged.) But I find this scenario fascinating, worthy of discussion, regardless of whether it is true or false of Jesuits. *If* the Jesuits practiced deliberate doubt, as described above, would they *therefore* be virtuous as rationalists?

I think I have to concede that the Jesuits, in the (possibly hypothetical) scenario above, would not properly be described as "fleeing from doubt". But the (possibly hypothetical) conduct still strikes me as highly suspicious. To a truly virtuous rationalist, doubt should not be scary. The conduct described above sounds to me like a program of desensitization for something *very* scary, like exposing an arachnophobe to spiders under carefully controlled conditions.

But even so, they are encouraging their novices to doubt—right? Does it matter if their reasons are flawed? Is this not still a worthy deed unto a rationalist?

All curiosity seeks to annihilate itself[3]; there is no curiosity that does not *want* an answer. But if you obtain an answer,

1. *http://lesswrong.com/lw/ib/the_proper_use_of_doubt/*
2. *http://yudkowsky.net/virtues/*
3. *http://yudkowsky.net/virtues/*

if you satisfy your curiosity, then the glorious mystery will no longer be mysterious.

In the same way, every doubt exists in order to annihilate some particular belief. If a doubt fails to destroy its target, the doubt has died unfulfilled—but that is still a resolution, an ending, albeit a sadder one. A doubt that neither destroys itself nor destroys its target might as well have never existed at all. It is the *resolution* of doubts, not the mere act of doubting, which drives the ratchet of rationality forward.

Every improvement is a change, but not every change is an improvement. Every rationalist doubts, but not all doubts are rational. Wearing doubts[4] doesn't make you a rationalist any more than wearing a white medical lab coat makes you a doctor.

A rational doubt comes into existence for a specific reason—you have some specific justification to suspect the belief is wrong. This reason in turn, implies an avenue of investigation which will either destroy the targeted belief, or destroy the doubt. This holds even for highly abstract doubts, like "I wonder if there might be a simpler hypothesis which also explains this data." In this case you investigate by trying to think of simpler hypotheses. As this search continues longer and longer without fruit, you will think it less and less likely that the next increment of computation will be the one to succeed. Eventually the cost of searching will exceed the expected benefit, and you'll stop searching. At which point you can no longer claim to be *usefully doubting*. A doubt that is not investigated might as well not exist. Every doubt exists to destroy itself, one way or the other. An unresolved doubt is a null-op; it does not turn the wheel, neither forward nor back.

If you really believe[5] a religion (not just believe in[6] it), then why would you tell your novices to consider doubts that must

4. Page 53, 'Belief as Attire'.
5. Page 39, 'Making Beliefs Pay Rent (in Anticipated Experiences)'.
6. Page 43, 'Belief in Belief'.

die unfulfilled? It would be like telling physics students to painstakingly doubt that the 20th-century revolution might have been a mistake, and that Newtonian mechanics was correct all along. If you don't *really* doubt something, why would you *pretend* that you do?

Because we all want to be seen as rational—and doubting is *widely believed* to be a virtue of a rationalist. But it is not widely understood that you need a particular reason to doubt, or that an unresolved doubt is a null-op. Instead people think it's about *modesty,* a submissive demeanor, maintaining the tribal status hierarchy—almost exactly the same problem as with humility, on which I have previously written[7]. Making a great public display of doubt to convince yourself[8] that you are a rationalist, will do around as much good as wearing a lab coat.

To avoid professing[9] doubts, remember:

- A rational doubt exists to destroy its target belief, and if it does not destroy its target it dies unfulfilled.
- A rational doubt arises from some specific reason the belief might be wrong.
- An unresolved doubt is a null-op.
- An uninvestigated doubt might as well not exist.
- You should not be proud of mere doubting, although you can justly be proud when you have just *finished* tearing a cherished belief to shreds.
- Though it may take courage to face your doubts, never forget that *to an ideal mind* doubt would not be scary in the first place.

7. Page 421, 'The Proper Use of Humility'.
8. Page 43, 'Belief in Belief'.
9. Page 50, 'Professing and Cheering'.

6. You Can Face Reality[1]

What is true is already so.
Owning up to it doesn't make it worse.
Not being open about it doesn't make it go away.
And because it's true, it is what is there to be interacted with.
Anything untrue isn't there to be lived.
People can stand what is true,
for they are already enduring it.
—Eugene Gendlin

(Hat tip to Stephen Omohundro.)

1. *http://lesswrong.com/lw/id/you_can_face_reality/*

7. The Meditation on Curiosity[1]

> "The first virtue is curiosity."
> —The *Twelve Virtues of Rationality*

As rationalists, we are obligated to criticize ourselves and question our beliefs... are we not?

Consider what happens to you, on a psychological level, if you begin by saying: "It is my duty to criticize my own beliefs." Roger Zelazny once distinguished between "wanting to be an author" versus "wanting to write". Mark Twain said: "A classic is something that everyone wants to have read and no one one wants to read." Criticizing yourself from a sense of duty leaves you *wanting to have investigated,* so that you'll be able to say afterward that your faith is not blind. This is not the same as *wanting to investigate.*

This can lead to motivated stopping[2] of your investigation. You consider an objection, then a counterargument to that objection, then you *stop there.* You repeat this with several objections, until you feel that you have done your duty to investigate, and then you *stop there.* You have achieved your underlying psychological objective: to get rid of the cognitive dissonance that would result from thinking of yourself as a rationalist, and yet knowing that you had not tried to criticize your belief. You might call it purchase of rationalist satisfaction[3]—trying to create a "warm glow" of discharged duty.

Afterward, your stated probability level will be high enough to justify your keeping the plans and beliefs you started with, but not so high as to evoke incredulity from yourself or other rationalists.

When you're really curious, you'll gravitate to inquiries that seem most promising of producing shifts in belief, or inquiries

1. *http://lesswrong.com/lw/jz/the_meditation_on_curiosity/*
2. Page 426, 'The Third Alternative'.
3. *http://lesswrong.com/lw/hw/scope_insensitivity/*

that are least like the ones you've tried before. Afterward, your probability distribution likely should *not* look like it did when you started out—shifts should have occurred, whether up or down; and either direction is equally fine to you, if you're genuinely curious.

Contrast this to the subconscious motive of keeping your inquiry on familiar ground, so that you can get your investigation over with quickly, so that you can *have investigated,* and restore the familiar balance on which your familiar old plans and beliefs are based.

As for what I think true curiosity should look like, and the power that it holds, I refer you to A Fable of Science and Politics[4]. Each of the characters is intended to illustrate different lessons. Ferris, the last character, embodies the power of innocent curiosity: which is lightness, and an eager reaching forth for evidence.

Ursula K. LeGuin wrote: "In innocence there is no strength against evil. But there is strength in it for good." Innocent curiosity may turn innocently awry; and so the training of a rationalist, and its accompanying sophistication,[5] must be dared as a danger if we want to become stronger[6]. Nonetheless we can try to keep the lightness and the eager reaching of innocence.

As it is written in the Twelve Virtues:

> "If in your heart you believe you already know, or if in your heart you do not wish to know, then your questioning will be purposeless and your skills without direction. Curiosity seeks to annihilate itself; there is no curiosity that does not want an answer."

4. Page 143, 'A Fable of Science and Politics'.
5. Page 333, 'Knowing About Biases Can Hurt People'.
6. *http://lesswrong.com/lw/h8/tsuyoku_naritai_i_want_to_become_stronger/*

There just isn't any good substitute for genuine curiosity. "A burning itch to know is higher than a solemn vow to pursue truth." But you can't produce curiosity just by willing it, any more than you can will your foot to feel warm when it feels cold. Sometimes, all we have is our mere solemn vows.

So what can you do with duty? For a start, we can try to take an interest in our dutiful investigations—keep a close eye out for sparks of genuine intrigue, or even genuine ignorance and a desire to resolve it. This goes right along with keeping a special eye out for possibilities that are painful,[7] that you are flinching away from—it's not all negative thinking.

It should also help to meditate on Conservation of Expected Evidence[8]. For every *new* point of inquiry, for every piece of *unseen* evidence that you suddenly look at, the expected posterior probability should equal your prior probability. In the microprocess of inquiry, your belief should always be evenly poised to shift in either direction. Not every point may suffice to blow the issue wide open—to shift belief from 70% to 30% probability—but if your current belief is 70%, you should be as ready to drop it to 69% as raising it to 71%. You should not think that you know which direction it will go in (on average), because by the laws of probability theory, if you know your destination, you are already there. If you can investigate honestly, so that each *new* point really does have equal potential to shift belief upward or downward, this may help to keep you interested or even curious about the microprocess of inquiry.

If the argument you are considering is *not* new, then why is your attention going here? Is this where you would look if you were genuinely curious? Are you subconsciously criticizing your belief at its strong points[9], rather than its weak points? Are you rehearsing the evidence[10]?

7. Page 357, 'Avoiding Your Belief's Real Weak Points'.
8. Page 68, 'Conservation of Expected Evidence'.
9. Page 357, 'Avoiding Your Belief's Real Weak Points'.
10. Page 340, 'One Argument Against An Army'.

If you can manage not to rehearse already known support, and you can manage to drop down your belief by one tiny bite at a time from the new evidence, you may even be able to relinquish the belief entirely—to realize from which quarter the winds of evidence are blowing against you.

Another restorative for curiosity is what I have taken to calling the Litany of Tarski, which is really a meta-litany that specializes for each instance (this is only appropriate). For example, if I am tensely wondering whether a locked box contains a diamond, then, rather than thinking about all the wonderful consequences if the box does contain a diamond, I can repeat the Litany of Tarski:

> *If the box contains a diamond,*
> *I desire to believe that the box contains a diamond;*
> *If the box does not contain a diamond,*
> *I desire to believe that the box does not contain a diamond;*
> *Let me not become attached to beliefs I may not want.*

Then you should meditate upon the possibility that there is no diamond, and the subsequent advantage that will come to you if you believe there is no diamond, and the subsequent disadvantage if you believe there is a diamond. See also the Litany of Gendlin[11].

If you can find within yourself the slightest shred of true uncertainty, then guard it like a forester nursing a campfire. If you can make it blaze up into a flame of curiosity, it will make you light and eager, and give purpose to your questioning and direction to your skills.

11. Page 477, 'You Can Face Reality'.

8. Something to Protect[1]

Followup to: Tsuyoku Naritai[2], Circular Altruism[3]

In the gestalt of (ahem) Japanese[4] fiction[5], one finds this oft-repeated motif: Power comes from having something to protect.

I'm not just talking about superheroes that power up when a friend is threatened, the way it works in Western fiction. In the Japanese version it runs deeper than that.

In the *X* saga it's explicitly stated that each of the good guys draw their power from having someone—one person—who they want to protect. Who? That question is part of *X*'s plot—the "most precious person" isn't always who we think. But if that person is killed, or hurt in the wrong way, the protector loses their power—not so much from magical backlash, as from simple despair. This isn't something that happens once per week per good guy, the way it would work in a Western comic. It's equivalent to being Killed Off For Real[6]—taken off the game board.

The way it works in Western superhero comics is that the good guy gets bitten by a radioactive spider; and then he needs something to do with his powers, to keep him busy, so he decides to fight crime. And then Western superheroes are always whining about how much time their superhero duties take up, and how they'd rather be ordinary mortals so they could go fishing or something.

Similarly, in Western real life, unhappy people are told that they need a "purpose in life", so they should pick out an altruistic cause that goes well with their personality, like picking out

1. *http://lesswrong.com/lw/nb/something_to_protect/*
2. *http://lesswrong.com/lw/h8/tsuyoku_naritai_i_want_to_become_stronger/*
3. *http://lesswrong.com/lw/n3/circular_altruism/*
4. *http://lesswrong.com/lw/m7/zen_and_the_art_of_rationality/*
5. Page 307, 'The Logical Fallacy of Generalization from Fictional Evidence'.
6. *http://tvtropes.org/pmwiki/pmwiki.php/Main/KilledOffForReal*

nice living-room drapes, and this will brighten up their days by adding some color, like nice living-room drapes. You should be careful not to pick something too expensive, though.

In Western comics, the magic comes first, then the purpose: Acquire amazing powers, decide to protect the innocent. In Japanese fiction, often, it works the other way around.

Of course I'm not saying all this to generalize from fictional evidence. But I want to convey a concept whose deceptively close Western analogue is *not* what I mean.

I have touched before on the idea that a rationalist must have something they value more than "rationality": *The Art must have a purpose other than itself, or it collapses into infinite recursion.* But do not mistake me, and think I am advocating that rationalists should pick out a nice altruistic cause, by way of having something to do, because rationality isn't all that important by itself. No. I am asking: Where do rationalists come from? How do we acquire our powers?

It is written in the *Twelve Virtues of Rationality:*

> How can you improve your conception of rationality? Not by saying to yourself, "It is my duty to be rational." By this you only enshrine your mistaken conception. Perhaps your conception of rationality is that it is rational to believe the words of the Great Teacher, and the Great Teacher says, "The sky is green," and you look up at the sky and see blue. If you think: "It may look like the sky is blue, but rationality is to believe the words of the Great Teacher," you lose a chance to discover your mistake.

Historically speaking, the way humanity *finally* left the trap of authority and began paying attention to, y'know, the actual sky, was that beliefs based on experiment turned out to be *much more useful* than beliefs based on authority. Curiosity has been around since the dawn of humanity, but the problem is that

spinning campfire tales works just as well[7] for satisfying curiosity.

Historically speaking, science won because it displayed greater raw strength in the form of technology, not because science *sounded more reasonable*. To this very day, magic and scripture still sound more reasonable to untrained ears than science. That is why there is continuous social tension between the belief systems. If science not only worked better than magic, but *also* sounded more intuitively reasonable, it would have won *entirely* by now.

Now there are those who say: "How dare you suggest that anything should be valued more than Truth? Must not a rationalist love Truth more than mere usefulness?"

Forget for a moment what would have happened historically to someone like that—that people in pretty much that frame of mind defended the Bible because they loved Truth more than mere accuracy. Propositional morality is a glorious thing, but it has too many degrees of freedom[8].

No, the real point is that a rationalist's love affair with the Truth is, well, just *more complicated* as an emotional relationship.

One doesn't become an adept rationalist without caring about the truth, both as a purely moral desideratum and as something that's fun to have. I doubt there are many master composers who hate music.

But part of what I *like* about rationality is the discipline imposed by requiring beliefs to yield predictions, which ends up taking us much closer to the truth than if we sat in the living room obsessing about Truth all day. I *like* the complexity of simultaneously having to love True-seeming ideas, and also being ready to drop them out the window at a moment's notice. I even like the glorious aesthetic purity of declaring that I value mere usefulness above aesthetics. That is almost a contradic-

7. Page 14, 'Why truth? And...'.
8. Page 14, 'Why truth? And...'.

tion, but not quite; and that has an aesthetic quality as well, a delicious humor.

And of course, no matter how much you profess your love of mere usefulness, you should never *actually* end up deliberately believing a useful false statement[9].

So don't oversimplify the relationship between loving truth and loving usefulness. It's not one or the other. It's *complicated,* which is not necessarily a defect in the moral aesthetics of single events[10].

But morality and aesthetics alone, believing that one ought to be "rational" or that certain ways of thinking are "beautiful", will not lead you to the center of the Way. It wouldn't have gotten humanity out of the authority-hole.

In Circular Altruism[11], I discussed this dilemma: Which of these options would you prefer:

1. Save 400 lives, with certainty
2. Save 500 lives, 90% probability; save no lives, 10% probability.

You may be tempted to grandstand, saying, "How dare you gamble with people's lives?" Even if you, yourself, are one of the 500—but you don't know which one—you may still be tempted to rely on the comforting feeling of certainty, because our own lives are often worth less to us than a good intuition[12].

But if your precious daughter is one of the 500, and you don't know which one, *then,* perhaps, you may feel more impelled to shut up and multiply—to notice that you have an 80% chance of saving her in the first case, and a 90% chance of saving her in the second.

And yes, everyone in that crowd is someone's son or daughter. Which, in turn, suggests that we should pick the second option as altruists, as well as concerned parents.

9. Page 401, 'Doublethink (Choosing to be Biased)'.
10. *http://lesswrong.com/lw/n9/the_intuitions_behind_utilitarianism/*
11. *http://lesswrong.com/lw/n3/circular_altruism/*
12. *http://lesswrong.com/lw/n9/the_intuitions_behind_utilitarianism/*

My point is not to suggest that one person's life is more valuable than 499 people. What I am trying to say is that *more* than your own life has to be at stake, before a person becomes desperate enough to resort to math.

What if you believe that it is "rational" to choose the certainty of option 1? Lots of people think that "rationality" is about choosing only methods that are certain to work, and rejecting all uncertainty. But, hopefully, you care more about your daughter's life than about "rationality".

Will pride in your own virtue as a rationalist save you? Not if you believe that it is virtuous to choose certainty. You will only be able to learn something about rationality if your daughter's life matters more to you than your pride as a rationalist.

You may even learn something about rationality from the experience, if you are already far enough grown in your Art to say, "I must have had the wrong conception of rationality," and not, "Look at how rationality gave me the wrong answer!"

(The essential difficulty in becoming a master rationalist is that you need quite a bit of rationality to bootstrap the learning process.)

Is your belief that you ought to be rational, more important than your life? Because, as I've previously observed, risking your life isn't comparatively all that scary. Being the lone voice of dissent[13] in the crowd and having everyone look at you funny is *much* scarier than a mere threat to your life, according to the revealed preferences of teenagers who drink at parties and then drive home. It will take something terribly important to make you willing to leave the pack. A threat to your life won't be enough.

Is your will to rationality stronger than your *pride?* Can it be, if your will to rationality stems from your pride in your self-image as a rationalist? It's helpful—*very* helpful—to have a self-image which says that you are the sort of person who confronts harsh truth. It's helpful to have too much self-respect to

13. Page 271, 'Lonely Dissent'.

knowingly lie to yourself or refuse to face evidence. But there may come a time when you have to admit that you've been doing rationality all wrong. Then your pride, your self-image as a rationalist, may make that too hard to face.

If you've prided yourself on believing what the Great Teacher says—even when it seems harsh, even when you'd rather not—that may make it all the more bitter a pill to swallow, to admit that the Great Teacher is a fraud, and all your noble self-sacrifice was for naught.

Where do you get the will to keep moving forward?

When I look back at my own personal journey toward rationality—not just humanity's historical journey—well, I grew up believing very strongly that I ought to be rational. This made me an above-average Traditional Rationalist a la Feynman and Heinlein, and nothing more. It did not drive me to go beyond the teachings I had received. I only began to grow *further* as a rationalist once I had something terribly important that I needed to do. Something more important than my pride as a rationalist, never mind my life.

Only when you become more wedded to success than to any of your beloved techniques of rationality, do you begin to appreciate these words of Miyamoto Musashi:

> "You can win with a long weapon, and yet you can also win with a short weapon. In short, the Way of the Ichi school is the spirit of winning, whatever the weapon and whatever its size."
>
> —Miyamoto Musashi, *The Book of Five Rings*

Don't mistake this for a specific teaching of rationality. It describes how you *learn* the Way, beginning with a desperate need to succeed. No one masters the Way until more than their life is at stake. More than their comfort, more even than their pride.

You can't just pick out a Cause[14] like that because you feel you need a hobby. Go looking for a "good cause", and your mind will just fill in a standard cliche[15]. Learn how to multiply[16], and perhaps you will recognize a drastically important cause when you see one.

But *if* you have a cause like that, it is right and proper to wield your rationality in its service.

To strictly subordinate the aesthetics of rationality to a higher cause, is part of the aesthetic of rationality. You should pay attention to that aesthetic: You will never master rationality well enough to win with any weapon, if you do not appreciate the beauty[17] for its own sake.

14. Page 247, 'Every Cause Wants To Be A Cult'.
15. Page 314, 'How to Seem (and Be) Deep'.
16. *http://lesswrong.com/lw/n9/the_intuitions_behind_utilitarianism/*
17. *http://lesswrong.com/lw/mt/beautiful_probability/*

9. No One Can Exempt You From Rationality's Laws[1]

Traditional Rationality is phrased in terms of *social rules,* with violations interpretable as cheating—as defections from cooperative norms. If you want me to accept a belief from you, you are obligated to provide me with a certain amount of evidence. If you try to get out of it, we all know you're cheating on your obligation. A theory is obligated to make bold predictions for itself, not just steal predictions that other theories have labored to make. A theory is obligated to expose itself to falsification—if it tries to duck out, that's like trying to duck out of a fearsome initiation ritual; you must pay your dues.

Traditional Rationality is phrased similarly to the customs that govern human societies, which makes it easy to pass on by word of mouth. Humans detect social cheating with much greater reliability than isomorphic violations of abstract logical rules. But viewing rationality as a social obligation gives rise to some strange ideas.

For example, one finds religious people defending their beliefs by saying, "Well, *you* can't justify your belief in science!" In other words, "How dare you criticize me for having unjustified beliefs, you hypocrite! You're doing it too!"

To Bayesians, the brain is an engine of accuracy: it processes and concentrates entangled evidence[2] into a map that reflects the territory[3]. The principles of rationality are laws[4] in the same sense as the second law of thermodynamics: obtaining a reliable belief requires a calculable amount of entangled evidence[5], just as reliably cooling the contents of a refrigerator requires a calculable minimum of free energy.

1. *http://lesswrong.com/lw/k1/no_one_can_exempt_you_from_rationalitys_laws/*
2. Page 18, 'What is Evidence?'.
3. *http://sl4.org/wiki/TheSimpleTruth*
4. *http://lesswrong.com/lw/hr/universal_law/*
5. Page 22, 'How Much Evidence Does It Take?'.

In principle, the laws of physics are time-reversible, so there's an infinitesimally tiny probability—indistinguishable from zero to all but mathematicians—that a refrigerator will spontaneously cool itself down while generating electricity. There's a slightly larger infinitesimal chance that you could accurately draw a detailed[6] street map of New York without ever visiting, sitting in your living room with your blinds closed and no Internet connection. But I wouldn't hold your breath.

Before you try mapping an unseen territory, pour some water into a cup at room temperature and wait until it spontaneously freezes before proceeding. That way you can be sure the general trick—ignoring infinitesimally tiny probabilities of success—is working properly. You might not realize directly that your map is wrong, especially if you never visit New York; but you can see that water doesn't freeze itself.

If the rules of rationality are social customs, then it may seem to excuse behavior X if you point out that others are doing the same thing. It wouldn't be *fair* to demand evidence from you, if we can't provide it ourselves. We will realize that none of us are better than the rest[7], and we will relent and mercifully excuse you from your social obligation to provide evidence for your belief. And we'll all live happily ever afterward in liberty, fraternity, and equality.

If the rules of rationality are mathematical laws, then trying to justify evidence-free belief by pointing to someone else doing the same thing, will be around as effective as listing 30 reasons why you shouldn't fall off a cliff. Even if we all vote that it's unfair for your refrigerator to need electricity, it still won't run (with probability 1). Even if we all vote that you shouldn't have to visit New York, the map will still be wrong. Lady Nature is famously indifferent to such pleading, and so is Lady Math.

So—to shift back to the social language of Traditional Rationality—don't think you can *get away with* claiming that it's

6. *http://lesswrong.com/lw/jk/burdensome_details/*
7. *http://lesswrong.com/lw/h9/tsuyoku_vs_the_egalitarian_instinct/*

okay to have arbitrary beliefs about XYZ, because other people have arbitrary beliefs too. If two parties to a contract both behave equally poorly, a human judge may decide to impose penalties on neither. But if two engineers design their engines equally poorly, neither engine will work. One design error cannot excuse another. Even if *I'm* doing XYZ wrong, it doesn't help you, or exempt you from the rules; it just means we're both screwed.

As a matter of human law in liberal democracies, everyone is entitled to their own beliefs. As a matter of Nature's law, you are not entitled to accuracy. We don't arrest people for believing weird things, at least not in the wiser countries. But no one can revoke the law[8] that you need evidence[9] to generate *accurate* beliefs[10]. Not even a vote of the whole human species can obtain mercy in the court of Nature.

Physicists don't decide the laws of physics, they just guess what they are. Rationalists don't decide the laws of rationality, we just guess what they are. You cannot "rationalize[11]" anything that is not rational to begin with. If by dint of extraordinary persuasiveness you convince all the physicists in the world that you are exempt from the law of gravity, and you walk off a cliff, you'll fall. Even saying "*We* don't decide" is too anthropomorphic. There is no higher authority that could exempt you. There is only cause and effect.

Remember this, when you plead to be excused just this once. We *can't* excuse you. It isn't up to us.

8. *http://lesswrong.com/lw/hr/universal_law/*
9. Page 18, 'What is Evidence?'.
10. *http://sl4.org/wiki/TheSimpleTruth*
11. Page 351, 'Rationalization'.

10. Leave a Line of Retreat[1]

"When you surround the enemy
Always allow them an escape route.
They must see that there is
An alternative to death."
—Sun Tzu, *The Art of War,* Cloud Hands edition

"Don't raise the pressure, lower the wall."
—Lois McMaster Bujold, *Komarr*

Last night I happened to be conversing with a nonrationalist who had somehow wandered into a local rationalists' gathering. She had just declared (a) her belief in souls and (b) that she didn't believe in cryonics because she believed the soul wouldn't stay with the frozen body. I asked, "But how do you know that?" From the confusion that flashed on her face, it was pretty clear that this question had never occurred to her. I don't say this in a bad way—she seemed like a nice person with absolutely no training in rationality, just like most of the rest of the human species. I really need to write that book.

Most of the ensuing conversation was on items already covered on Overcoming Bias—if you're *really* curious about something, you probably *can* figure out a good way to test it; try to attain accurate beliefs first and then let your emotions flow from that—that sort of thing. But the conversation reminded me of one notion I haven't covered here yet:

"Make sure," I suggested to her, "that you visualize what the world would be like if there are no souls, and what you would do about that. Don't think about all the reasons that it can't be that way, just accept it as a premise and then visualize the consequences. So that you'll think, 'Well, if there are no souls, I can just sign up for cryonics', or 'If there is no God, I can just go on being moral anyway,' rather than it being too horrifying to face.

1. *http://lesswrong.com/lw/o4/leave_a_line_of_retreat/*

As a matter of self-respect you should try to believe the truth no matter how uncomfortable it is, like I said before; but as a matter of human nature, it helps to make a belief less uncomfortable, *before* you try to evaluate the evidence for it."

The principle behind the technique is simple: As Sun Tzu advises you to do with your enemies, you must do with yourself—leave yourself a line of retreat, so that you will have less trouble retreating. The prospect of losing your job, say, may seem a lot more scary when you can't even bear to think about it, than after you have calculated exactly how long your savings will last, and checked the job market in your area, and otherwise planned out exactly what to do next. Only then will you be ready to *fairly* assess the probability of keeping your job in the planned layoffs next month. Be a true coward, and plan out your retreat in detail—visualize every step—preferably before you first come to the battlefield.

The hope is that it takes less courage to visualize an uncomfortable state of affairs *as a thought experiment*, than to consider *how likely* it is to be true. But then after you do the former, it becomes easier to do the latter.

Remember that Bayesianism is precise—even if a scary proposition really should seem unlikely, it's still important to count up all the evidence, for and against, exactly fairly, to arrive at the rational quantitative probability. Visualizing a scary belief does *not* mean admitting that you think, deep down, it's probably true. You can visualize a scary belief on general principles of good mental housekeeping. "The thought you cannot think controls you more than thoughts you speak aloud"—this happens even if the unthinkable thought is false!

The leave-a-line-of-retreat technique does require a certain minimum of self-honesty to use correctly.

For a start: You must at least be able to admit to yourself *which* ideas scare you, and which ideas you are attached to. But this is a substantially less difficult test than fairly counting the evidence for an idea that scares you. Does it help if I say that I

have occasion to use this technique myself? A rationalist does not reject all emotion, after all. There are ideas which scare me, yet I still believe to be false. There are ideas to which I know I am attached, yet I still believe to be true. But I still plan my retreats, not because I'm planning *to* retreat, but because planning my retreat in advance helps me think about the problem without attachment.

But greater test of self-honesty is to *really* accept the uncomfortable proposition as a premise, and figure out how you would *really* deal with it. When we're faced with an uncomfortable idea, our first impulse is naturally to think of all the reasons why it *can't possibly* be so. And so you will encounter a certain amount of psychological resistance in yourself, if you try to visualize exactly how the world would be, and what you would do about it, if My-Most-Precious-Belief were false, or My-Most-Feared-Belief were true.

Think of all the people who say that, without God, morality was impossible. (And yes, this topic did come up in the conversation; so I am not offering a strawman.) If theists could visualize their *real* reaction to believing as a fact that God did not exist, they could realize that, no, they wouldn't go around slaughtering babies. They could realize that atheists are reacting to the nonexistence of God in pretty much the way they themselves would, if they came to believe that. I say this, to show that it *is* a considerable challenge to visualize the way you *really would* react, to believing the opposite of a tightly held belief.

Plus it's always counterintuitive to realize that, yes, people do get over things. Newly minted quadriplegics are not as sad as they expect to be six months later, etc. It can be equally counterintuitive to realize that if the scary belief turned out to be true, you *would* come to terms with it somehow. Quadriplegics deal, and so would you.

See also the Litany of Gendlin[2] and the Litany of Tarski[3]. What is true is already so; owning up to it doesn't make it worse. You shouldn't be afraid to just *visualize* a world you fear. If that world is already actual, visualizing it won't make it worse; and if it is *not* actual, visualizing it will do no harm. And remember, as you visualize, that if the scary things you're imagining really are true—which they may not be!—then you would, indeed, want to believe it, and you should visualize that too; not believing wouldn't help you.

How many religious people would retain their belief in God, if they could *accurately* visualize that hypothetical world in which there was no God and they themselves have become atheists?

Leaving a line of retreat is a powerful technique, but it's not easy. *Honest* visualization doesn't take as much effort as admitting *outright* that God doesn't exist, but it does take an effort.

> (*Meta note:* I'm posting this on the advice that I should break up long sequences of mathy posts with non-mathy posts. (I was actually advised to post something "fun", but I'd rather not—it feels like I have too much important material to cover in the next couple of months.) If anyone thinks that I should have, instead, gone ahead and posted the next item in the information-theory sequence rather than breaking it up; or, alternatively, thinks that this non-mathy post came as a welcome change; then I am interested in hearing from you in the comments.)

2. Page 357, 'Avoiding Your Belief's Real Weak Points'.
3. Page 478, 'The Meditation on Curiosity'.

11. Crisis of Faith[1]

Followup to: Make an Extraordinary Effort[2], The Meditation on Curiosity[3], Avoiding Your Belief's Real Weak Points[4]

> "It ain't a true crisis of faith unless things could just as easily go either way."
>
> —Thor Shenkel

Many in this world retain beliefs whose flaws a ten-year-old could point out, *if* that ten-year-old were hearing the beliefs for the first time. These are not subtle errors we are talking about. They would be child's play for an unattached[5] mind to relinquish, if the skepticism of a ten-year-old were applied without evasion[6]. As Premise Checker put it, "Had the idea of god not come along until the scientific age, only an exceptionally weird person would invent such an idea and pretend that it explained[7] anything."

And yet skillful scientific specialists, even the major innovators of a field, even in this very day and age, do not apply that skepticism successfully. Nobel laureate Robert Aumann, of Aumann's Agreement Theorem, is an Orthodox Jew: I feel reasonably confident in venturing that Aumann must, at one point or another, have questioned his faith. And yet he did not doubt successfully[8]. We change our minds less often than we think.[9]

---

1. *http://lesswrong.com/lw/ur/crisis_of_faith/*
2. *http://lesswrong.com/lw/uo/make_an_extraordinary_effort/*
3. Page 478, 'The Meditation on Curiosity'.
4. Page 357, 'Avoiding Your Belief's Real Weak Points'.
5. Page 225, 'Affective Death Spirals'.
6. Page 357, 'Avoiding Your Belief's Real Weak Points'.
7. Page 29, 'Occam's Razor'.
8. Page 474, 'The Proper Use of Doubt'.
9. Page 318, 'We Change Our Minds Less Often Than We Think'.

This should scare you down to the marrow of your bones. It means you can be a world-class scientist *and* conversant with Bayesian mathematics *and* still fail to reject a belief whose absurdity a fresh-eyed ten-year-old could see. It shows the invincible defensive position which a belief can create for itself, if it has long festered in your mind.

What does it take to defeat an error which has built itself a fortress?

But by the time you *know* it is an error, it is already defeated. The dilemma is not "How can I reject long-held false belief X?" but "How do I know if long-held belief X is false?" Self-honesty is at its most fragile when we're not *sure* which path is the righteous one. And so the question becomes:

How can we create in ourselves a true crisis of faith, that could just as easily go either way?

Religion is the trial case we can all imagine. (Readers born to atheist parents have missed out on a fundamental life trial, and must make do with the poor substitute of thinking of their religious friends.) But if you have cut off all sympathy and now think of theists as evil mutants[10], then you won't be able to imagine the real internal trials they face. You won't be able to ask the question:

"What general strategy would a religious person have to follow in order to escape their religion?"

I'm sure that some, looking at this challenge, are already rattling off a list of standard atheist talking points—"They would have to admit that there wasn't any Bayesian evidence for God's existence", "They would have to see the moral evasions they were carrying out to excuse God's behavior in the Bible", "They need to learn how to use Occam's Razor—"

WRONG! WRONG WRONG WRONG! This kind of rehearsal[11], where you just cough up points *you already thought of long before,* is *exactly* the style of thinking that keeps people

10. Page 160, 'Are Your Enemies Innately Evil?'.
11. Page 340, 'One Argument Against An Army'.

within their current religions. If you stay with your cached thoughts[12], if your brain fills in the obvious answer so fast that you can't see originally[13], you surely will not be able to conduct a crisis of faith.

Even when it's explicitly pointed out, some people seemingly *cannot follow the leap* from the object-level "Use Occam's Razor! You have to see that your God is an unnecessary belief!" to the meta-level "Try to stop your mind from completing the pattern the usual way!" Because in the same way that all your rationalist friends talk about Occam's Razor like it's a good thing, and in the same way that Occam's Razor leaps right up into your mind, so too, the obvious friend-approved religious response is "God's ways are mysterious and it is presumptuous to suppose that we can understand them." So for you to think that the *general* strategy to follow is "Use Occam's Razor", would be like a theist saying that the general strategy is to have faith. (I've noticed that a large fraction of the population—even technical folk—have trouble following [14]arguments that go this meta. On my more pessimistic days I wonder if the camel has two humps[15].)

"But—but Occam's Razor really is better than faith! That's not like preferring a different flavor of ice cream! Anyone can see, looking at history, that Occamian reasoning has been far more productive than faith—"

Which is all true. But beside the point. The point is that you, saying this, are rattling off a standard justification that's already in your mind. The challenge of a crisis of faith is to handle the case where, possibly, our standard conclusions are *wrong* and our standard justifications are *wrong*. So if the standard justification for X is "Occam's Razor!", and you want to hold a crisis of faith around X, you should be questioning

12. Page 297, 'Cached Thoughts'.
13. Page 304, 'Original Seeing'.
14. *http://lesswrong.com/lw/h6/chronophone_motivations/*
15. *https://www.cs.kent.ac.uk/dept_info/seminars/2005_06/paper1.pdf*

if Occam's Razor really endorses X, if your understanding of Occam's Razor is correct, and—if you want to have sufficiently deep doubts—whether simplicity *is* the sort of criterion that has worked well historically in this case, or could reasonably be *expected* to work, etcetera. If you would advise a religionist to question their belief that "faith" is a good justification for X, then you should advise yourself to put forth an equally strong effort to question your belief that "Occam's Razor" is a good justification for X.

(Think of all the people out there who don't understand the Minimum Description Length or Solomonoff Induction formulations of Occam's Razor, who think that Occam's Razor outlaws Many-Worlds[16] or the Simulation Hypothesis[17]. They would need to question their formulations of Occam's Razor and their notions of why simplicity is a good thing. Whatever X in contention you just justified by saying "Occam's Razor!", I bet it's not the same level of Occamian slam dunk as gravity.)

If "Occam's Razor!" is your usual reply, your standard reply, the reply that all your friends give—then you'd better block your brain from instantly completing that pattern, if you're trying to instigate a true crisis of faith.

Better to think of such rules as, "Imagine what a skeptic would say—and then imagine what they would say to your response—and then imagine what else they might say, that would be harder to answer."

Or, "Try to think the thought that hurts the most."

And above all, the rule:

"Put forth the same level of desperate effort[18] that it would take for a theist to reject their religion."

Because, if you *aren't* trying that hard, then—for all *you* know—your head could be stuffed full of nonsense as ridiculous as religion.

16. *http://lesswrong.com/lw/q3/decoherence_is_simple/*
17. *http://www.simulation-argument.com/simulation.html*
18. *http://lesswrong.com/lw/uo/make_an_extraordinary_effort/*

Without a convulsive, wrenching effort to be rational, the kind of effort it would take to throw off a religion—then how dare you believe anything, when Robert Aumann believes in God?

Someone (I forget who) once observed that people had only until a certain age to reject their religious faith. Afterward they would have answers to all the objections, and it would be too late. That is the kind of existence you must surpass. This is a test of your strength as a rationalist, and it is very severe; but if you cannot pass it, you will be weaker than a ten-year-old.

But again, by the time you know a belief is an error, it is already defeated. So we're not talking about a desperate, convulsive effort to undo the effects[19] of a religious upbringing, *after* you've come to the conclusion that your religion is wrong. We're talking about a desperate effort to *figure out* if you should be throwing off the chains, or keeping them. Self-honesty is at its most fragile when we don't *know* which path we're supposed to take—that's when rationalizations are not *obviously* sins.

Not every doubt calls for staging an all-out Crisis of Faith. But you should consider it when:

- A belief has long remained in your mind;
- It is surrounded by a cloud of known arguments and refutations;
- You have sunk costs[20] in it (time, money, public declarations);
- The belief has emotional consequences[21] (note this does not make it wrong);
- It has gotten mixed up in your personality generally.

None of these warning signs are immediate disproofs. These attributes place a belief at-risk[22] for all sorts of dangers,

19. Page 327, 'The Genetic Fallacy'.
20. *http://en.wikipedia.org/wiki/Sunk_cost*
21. Page 463, 'Feeling Rational'.
22. Page 276, 'Cultish Countercultishness'.

and make it very hard to reject when it *is* wrong. But they also hold for Richard Dawkins's belief in evolutionary biology as well as the Pope's Catholicism. This does not say that we are only talking about different flavors of ice cream. Only the unenlightened think that all deeply-held beliefs are on the same level regardless of the evidence supporting them, just because they are deeply held. The point is not to have shallow beliefs, but to have a map which reflects the territory.

I emphasize this, of course, so that you can admit to yourself, "My belief has these warning signs," without having to say to yourself, "My belief is false."

But what these warning signs *do* mark, is a belief that will take *more than an ordinary effort to doubt effectively*. So that if it were in fact false, you would in fact reject it. And where you cannot doubt effectively, you are blind, because your brain will hold the belief unconditionally[23]. When a retina sends the same signal regardless of the photons entering it, we call that eye blind[24].

When should you stage a Crisis of Faith?

Again, think of the advice you would give to a theist: If you find yourself feeling a little unstable inwardly, but trying to rationalize reasons the belief is still solid, then you should probably stage a Crisis of Faith. If the belief is as solidly supported as gravity, you needn't bother—but think of all the theists who would desperately want to conclude that God is as solid as gravity. So try to imagine what the skeptics out there would say to your "solid as gravity" argument. Certainly, one reason you might fail at a crisis of faith is that you never really sit down and question in the first place—that you never say, "Here is something I need to put effort into doubting properly."

If your thoughts get that complicated, you should go ahead and stage a Crisis of Faith. Don't try to do it haphazardly, don't try it in an ad-hoc spare moment. Don't rush to get it done with

23. Page 343, 'The Bottom Line'.
24. Page 18, 'What is Evidence?'.

quickly, so that you can say "I have doubted as I was obliged to do." That wouldn't work for a theist and it won't work for you either. Rest up the previous day, so you're in good mental condition. Allocate some uninterrupted hours. Find somewhere quiet to sit down. Clear your mind of all standard arguments, try to see from scratch. And make a desperate effort to put forth a true doubt that would destroy a false, and *only* a false, deeply held belief.

Elements of the Crisis of Faith technique have been scattered over many posts:

- Avoiding Your Belief's Real Weak Points[25]—One of the first temptations in a crisis of faith is to doubt the strongest points of your belief, so that you can rehearse[26] your good answers. You need to seek out the most painful spots, not the arguments that are most reassuring to consider.
- The Meditation on Curiosity[27]—Roger Zelazny once distinguished between "wanting to be an author" versus "wanting to write", and there is likewise a distinction between wanting to have investigated and wanting to investigate. It is not enough to say "It is my duty to criticize my own beliefs"; you must be curious, and only uncertainty can create curiosity. Keeping in mind Conservation of Expected Evidence[28] may help you Update Yourself Incrementally[29]: For every *single* point that you consider, and each element of new argument and new evidence, you should not expect your beliefs to shift more (on average) in one direction than another—thus you can be truly curious each time about how it will go.

25. Page 357, 'Avoiding Your Belief's Real Weak Points'.
26. Page 340, 'One Argument Against An Army'.
27. Page 478, 'The Meditation on Curiosity'.
28. Page 68, 'Conservation of Expected Evidence'.
29. Page 336, 'Update Yourself Incrementally'.

- Cached Thoughts[30] and Pirsig's Original Seeing[31], to prevent standard thoughts from rushing in and completing the pattern.
- The Litany of Gendlin[32] and the Litany of Tarski[33]: People can stand what is true, for they are already enduring it. If a belief is true you will be better off believing it, and if it is false you will be better off rejecting it. You would advise a religious person to try to visualize fully and deeply the world in which there is no God, and to, without excuses, come to the full understanding that *if* there is no God *then* they will be better off believing there is no God. If one cannot come to accept this on a deep emotional level, they will not be able to have a crisis of faith. So you should put in a sincere effort to visualize the *alternative* to your belief, the way that the best and highest skeptic would want you to visualize it. Think of the effort a religionist would have to put forth to imagine, without corrupting it for their own comfort, an atheist's view of the universe.
- Make an Extraordinary Effort[34], for the concept of *isshokenmei*, the desperate convulsive effort to be rational that it would take to surpass the level of Robert Aumann and all the great scientists throughout history who never let go of their religions.
- The Genetic Heuristic[35]: You should be extremely suspicious if you have many ideas suggested by a source that you now know to be untrustworthy, but by golly, it seems that all the ideas still ended up being right. (E.g., the one concedes that the Bible was

30. Page 297, 'Cached Thoughts'.
31. Page 304, 'Original Seeing'.
32. Page 477, 'You Can Face Reality'.
33. Page 478, 'The Meditation on Curiosity'.
34. *http://lesswrong.com/lw/uo/make_an_extraordinary_effort/*
35. Page 327, 'The Genetic Fallacy'.

written by human hands, but still clings to the idea that it contains indispensable ethical wisdom[36].)

- The Importance of Saying "Oops"[37]—it really is less painful to swallow the entire bitter pill in one terrible gulp.
- Singlethink[38], the opposite of doublethink. See the thoughts you flinch away from, that appear in the corner of your mind for just a moment [39]before you refuse to think them. If you become aware of what you are not thinking, you can think it.
- Affective Death Spirals[40] and Resist the Happy Death Spiral[41]. Affective death spirals are prime generators of false beliefs that it will take a Crisis of Faith to shake loose. But since affective death spirals can also get started around real things that are genuinely nice, you don't have to admit that your belief is a lie, to try and resist the halo effect at every point—refuse false praise even of genuinely nice things. Policy debates should not appear one-sided[42].
- Hold Off On Proposing Solutions[43] until the problem has been discussed as thoroughly as possible without proposing any; make your mind hold off from knowing what its answer will be[44]; and try for five minutes before giving up[45], both generally, and especially when pursuing the devil's point of view.

And these standard techniques are particularly relevant:

36. *http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/*
37. Page 466, 'The Importance of Saying "Oops"'.
38. Page 399, 'Singlethink'.
39. Page 399, 'Singlethink'.
40. Page 225, 'Affective Death Spirals'.
41. Page 228, 'Resist the Happy Death Spiral'.
42. Page 150, 'Policy Debates Should Not Appear One-Sided'.
43. Page 320, 'Hold Off On Proposing Solutions'.
44. Page 318, 'We Change Our Minds Less Often Than We Think'.
45. *http://lesswrong.com/lw/ui/use_the_try_harder_luke/*

- The sequence on The Bottom Line[46] and Rationalization[47], which explains why it is always wrong to selectively argue one side of a debate.
- Positive Bias[48] and motivated skepticism[49] and motivated stopping[50], lest you selectively look for support, selectively look for counter-counterarguments, and selectively stop the argument before it gets dangerous. Missing alternatives[51] are a special case of stopping. A special case of motivated skepticism is fake humility[52] where you bashfully confess that no one can know[53] something you would rather not know. Don't selectively demand too much authority[54] of counterarguments.
- Beware of Semantic Stopsigns[55], Applause Lights[56], and the choice to Explain/Worship/Ignore[57].
- Feel the weight of Burdensome Details[58]; each detail a separate burden, a point of crisis.

But really there's rather a lot of relevant material, here and there on *Overcoming Bias*. The Crisis of Faith is only the critical point and sudden clash of the longer *isshoukenmei*—the lifelong uncompromising effort to be so incredibly rational that you rise above the level of stupid damn mistakes. It's when you

46. Page 343, 'The Bottom Line'.
47. Page 351, 'Rationalization'.
48. Page 108, 'Positive Bias: Look Into the Dark'.
49. Page 333, 'Knowing About Biases Can Hurt People'.
50. Page 362, 'Motivated Stopping and Motivated Continuation'.
51. Page 426, 'The Third Alternative'.
52. Page 421, 'The Proper Use of Humility'.
53. *http://lesswrong.com/lw/kj/no_one_knows_what_science_doesnt_know/*
54. Page 443, 'Absolute Authority'.
55. Page 92, 'Semantic Stopsigns'.
56. Page 128, 'Applause Lights'.
57. Page 122, 'Explain/Worship/Ignore?'.
58. *http://lesswrong.com/lw/jk/burdensome_details/*

get a chance to use your skills that you've been practicing for so long, all-out against yourself.

I wish you the best of luck against your opponent. Have a wonderful crisis!

12. The Ritual[1]

Followup to: The Failures of Eld Science[2], Crisis of Faith[3]

The room in which Jeffreyssai received his non-*beisutsukai* visitors was quietly formal, impeccably appointed in only the most conservative tastes. Sunlight and outside air streamed through a grillwork of polished silver, a few sharp edges making it clear that this wall was not to be opened. The floor and walls were glass, thick enough to distort, to a depth sufficient that it didn't matter what might be underneath. Upon the surfaces of the glass were subtly scratched patterns of no particular meaning, scribed as if by the hand of an artistically inclined child (and this was in fact the case).

Elsewhere in Jeffreyssai's home there were rooms of other style; but this, he had found, was what most outsiders expected of a Bayesian Master, and he chose not to enlighten them otherwise. That quiet amusement was one of life's little joys, after all.

The guest sat across from him, knees on the pillow and heels behind. She was here solely upon the business of her Conspiracy, and her attire showed it: A form-fitting jumpsuit of pink leather with even her hands gloved—all the way to the hood covering her head and hair, though her face lay plain and unconcealed beneath.

And so Jeffreyssai had chosen to receive her in this room.

Jeffreyssai let out a long breath, exhaling. "Are you *sure?*"

"Oh," she said, "and do I have to be *absolutely certain* before my advice can shift your opinions? Does it not suffice that I am a domain expert, and you are not?"

Jeffreyssai's mouth twisted up at the corner in a half-smile. "How do *you* know so much about the rules, anyway? You've never had so much as a Planck length of formal training."

1. *http://lesswrong.com/lw/us/the_ritual/*
2. *http://lesswrong.com/lw/q9/the_failures_of_eld_science/*
3. Page 496, 'Crisis of Faith'.

"Do you even need to ask?" she said dryly. "If there's one thing that you *beisutsukai* do love to go on about, it's the reasons why you do things."

Jeffreyssai inwardly winced at the thought of trying to pick up rationality by watching other people talk about it—

"And don't inwardly wince at me like that," she said. "I'm not trying to be a rationalist myself, just trying to win an argument with a rationalist. There's a difference, as I'm sure you tell your students."

Can she really read me that well? Jeffreyssai looked out through the silver grillwork, at the sunlight reflected from the faceted mountainside. Always, always the golden sunlight fell each day, in this place far above the clouds. An unchanging thing, that light. The distant Sun, which that light represented, was in five billion years burned out; but now, in *this* moment, the Sun still shone. And that could never alter. Why wish for things to stay the same way forever, when that wish was already granted as absolutely as any wish could be? The paradox of permanence and impermanence: only in the latter perspective was there any such thing as progress, or loss.

"You have always given me good counsel," Jeffreyssai said. "Unchanging, that has been. Through all the time we've known each other."

She inclined her head, acknowledging. This was true, and there was no need to spell out the implications.

"So," Jeffreyssai said. "Not for the sake of arguing. Only because I want to know the answer. *Are* you sure?" He didn't even see how she could *guess*.

"Pretty sure," she said, "we've been collecting statistics for a long time, and in nine hundred and eight-five out of a thousand cases like yours—"

Then she laughed at the look on his face. "No, I'm joking. Of course I'm not sure. This thing only you can decide. But I *am* sure that you should go off and do whatever it is you people do—I'm quite sure you have a ritual for it, even if you

won't discuss it with outsiders—when you *very seriously consider* abandoning a long-held premise of your existence."

It was hard to argue with that, Jeffreyssai reflected, the more so when a domain expert had told you that you were, in fact, probably wrong.

"I concede," Jeffreyssai said. Coming from his lips, the phrase was spoken with a commanding finality. *There is no need to argue with me any further: You have won.*

"Oh, stop it," she said. She rose from her pillow in a single fluid shift without the slightest wasted motion. She didn't flaunt her age, but she didn't conceal it either. She took his outstretched hand, and raised it to her lips for a formal kiss. "Farewell, sensei."

"Farewell?" repeated Jeffreyssai. That signified a higher order of departure than *goodbye*. "I do intend to visit you again, milady; and you are always welcome here."

She walked toward the door without answering. At the doorway she paused, without turning around. "It won't be the same," she said. And then, without the movements seeming the least rushed, she walked away so swiftly it was almost like vanishing.

Jeffreyssai sighed. But at least, from here until the challenge proper, all his actions were prescribed, known quantities.

Leaving that formal reception area, he passed to his arena, and caused to be sent out messengers to his students, telling them that the next day's classes must be improvised in his absence, and that there would be a test later.

And then he did nothing in particular. He read another hundred pages of the textbook he had borrowed; it wasn't very good, but then the book he had loaned out in exchange wasn't very good either. He wandered from room to room of his house, idly checking various storages to see if anything had been stolen (a deck of cards was missing, but that was all). From time to time his thoughts turned to tomorrow's challenge, and he let them drift. Not directing his thoughts at all, only

blocking out every thought that had ever *previously* occurred to him; and disallowing any kind of conclusion, or even any thought as to where his thoughts might be trending.

The sun set, and he watched it for a while, mind carefully put in idle. It was a fantastic balancing act to set your mind in idle without having to obsess about it, or exert energy to keep it that way; and years ago he would have sweated over it, but practice had long since made perfect.

The next morning he awoke with the chaos of the night's dreaming fresh in his mind, and, doing his best to preserve the feeling of the chaos as well as its memory, he descended a flight of stairs, then another flight of stairs, then a flight of stairs after that, and finally came to the least fashionable room in his whole house.

It was white. That was pretty much it as far as the color scheme went.

All along a single wall were plaques, which, following the classic and suggested method, a younger Jeffreyssai had very carefully scribed himself, burning the *concepts* into his mind with each touch of the brush that wrote the words. *That which can be destroyed by the truth should be. People can stand what is true, for they are already enduring it. Curiosity seeks to annihilate itself.* Even one small plaque that showed nothing except a red horizontal slash. Symbols could be made to stand for *anything;* a flexibility of visual power that even the Bardic Conspiracy would balk at admitting outright.

Beneath the plaques, two sets of tally marks scratched into the wall. Under the plus column, two marks. Under the minus column, five marks. Seven times he had entered this room; five times he had decided not to change his mind; twice he had exited something of a different person. There was no set ratio prescribed, or set range—that would have been a mockery indeed. But if there were no marks in the plus column after a while, you might as well admit that there was no point in having the room, since you didn't have the ability it stood for. Either

that, or you'd been born knowing the truth and right of everything.

Jeffreyssai seated himself, not facing the plaques, but facing away from them, at the featureless white wall. It was better to have no visual distractions.

In his mind, he rehearsed first the meta-mnemonic, and then the various sub-mnemonics referenced, for the seven major principles and sixty-two specific techniques that were most likely to prove needful in the Ritual Of Changing One's Mind. To this, Jeffreyssai added another mnemonic, reminding himself of his own fourteen most embarrassing oversights.

He did not take a deep breath. Regular breathing was best.

And then he asked himself the question.