



# LessWrong.com Sequences

Elizier Yudkowsky



## **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.*



## I. The Simple Truth<sup>1</sup>

“I remember this paper I wrote on existentialism. My teacher gave it back with an F. She’d underlined true and truth wherever it appeared in the essay, probably about twenty times, with a question mark beside each. She wanted to know what I meant by truth.”

— Danielle Egan (journalist)

### *Author’s Foreword:*

This essay is meant to restore a naive view of truth.

Someone says to you: “My miracle snake oil can rid you of lung cancer in just three weeks.” You reply: “Didn’t a clinical study show this claim to be untrue?” The one returns: “This notion of ‘truth’ is quite naive; what do you mean by ‘true’?”

Many people, so questioned, don’t know how to answer in exquisitely rigorous detail. Nonetheless they would not be wise to abandon the concept of ‘truth’. There was a time when no one knew the equations of gravity in exquisitely rigorous detail, yet if you walked off a cliff, you would fall.

Often I have seen – especially on Internet mailing lists – that amidst other conversation, someone says “X is true”, and then an argument breaks out over the use of the word ‘true’. This essay is *not* meant as an encyclopedic reference for that argument. Rather, I hope the arguers will read this essay, and then go back to whatever they were discussing before someone questioned the nature of truth.

In this essay I pose questions. If you see what seems like a really obvious answer, it’s probably the answer I intend. The obvious choice isn’t *always* the best choice, but sometimes, by golly, it *is*. I don’t stop looking as soon I find an obvious answer, but if I go on looking, and the obvious-seeming answer *still* seems obvious, I don’t feel guilty about keeping it. Oh, sure, everyone *thinks* two plus two is four, everyone *says* two plus two is four, and in the mere mundane drudgery of everyday life everyone *behaves* as if two plus two is four, but what does two plus two *really, ultimately* equal? As near as I can figure, four. It’s still four even if I intone the question in a

---

<sup>1</sup>. <http://yudkowsky.net/rational/the-simple-truth>

solemn, portentous tone of voice. Too simple, you say? Maybe, on this occasion, life doesn't *need* to be complicated. Wouldn't that be refreshing?

If you are one of those fortunate folk to whom the question seems trivial at the outset, I hope it still seems trivial at the finish. If you find yourself stumped by deep and meaningful questions, remember that if you know exactly how a system works, and could build one yourself out of buckets and pebbles, it should not be a mystery to you.

If confusion threatens when you interpret a metaphor as a metaphor, try taking everything *completely literally*.

---

Imagine that in an era before recorded history or formal mathematics, I am a shepherd and I have trouble tracking my sheep. My sheep sleep in an enclosure, a fold; and the enclosure is high enough to guard my sheep from wolves that roam by night. Each day I must release my sheep from the fold to pasture and graze; each night I must find my sheep and return them to the fold. If a sheep is left outside, I will find its body the next morning, killed and half-eaten by wolves. But it is so discouraging, to scour the fields for hours, looking for one last sheep, when I know that probably all the sheep are in the fold. Sometimes I give up early, and usually I get away with it; but around a tenth of the time there is a dead sheep the next morning.

If only there were some way to divine whether sheep are still grazing, without the inconvenience of looking! I try several methods: I toss the divination sticks of my tribe; I train my psychic powers to locate sheep through clairvoyance; I search carefully for reasons to believe all the sheep are in the fold. It makes no difference. Around a tenth of the times I turn in early, I find a dead sheep the next morning. Perhaps I realize that my methods aren't working, and perhaps I carefully excuse each failure; but my dilemma is still the same. I can spend an hour searching every possible nook and cranny, when most of the time there are no remaining sheep; or I can go to sleep early and lose, on the average, one-tenth of a sheep.

Late one afternoon I feel especially tired. I toss the divination sticks and the divination sticks say that all the sheep have returned. I visualize each nook and cranny, and I don't imagine scrying any

sheep. I'm still not confident enough, so I look inside the fold and it seems like there are a lot of sheep, and I review my earlier efforts and decide that I was especially diligent. This dissipates my anxiety, and I go to sleep. The next morning I discover *two* dead sheep. Something inside me snaps, and I begin thinking creatively.

That day, loud hammering noises come from the gate of the sheepfold's enclosure.

The next morning, I open the gate of the enclosure only a little way, and as each sheep passes out of the enclosure, I drop a pebble into a bucket nailed up next to the door. In the afternoon, as each returning sheep passes by, I take one pebble out of the bucket. When there are no pebbles left in the bucket, I can stop searching and turn in for the night. It is a *brilliant* notion. It will revolutionize shepherding.

That was the theory. In practice, it took considerable refinement before the method worked reliably. Several times I searched for hours and didn't find any sheep, and the next morning there were no stragglers. On each of these occasions it required deep thought to figure out where my bucket system had failed. On returning from one fruitless search, I thought back and realized that the bucket already contained pebbles when I started; this, it turned out, was a bad idea. Another time I randomly tossed pebbles into the bucket, to amuse myself, between the morning and the afternoon; this too was a bad idea, as I realized after searching for a few hours. But I practiced my pebblecraft, and became a reasonably proficient pebblecrafter.

One afternoon, a man richly attired in white robes, leafy laurels, sandals, and business suit trudges in along the sandy trail that leads to my pastures.

"Can I help you?" I inquire.

The man takes a badge from his coat and flips it open, proving beyond the shadow of a doubt that he is Markos Sophisticus Maximus, a delegate from the Senate of Rum. (One might wonder whether another could steal the badge; but so great is the power of these badges that if any other were to use them, they would in that instant be *transformed* into Markos.)

"Call me Mark," he says. "I'm here to confiscate the magic pebbles, in the name of the Senate; artifacts of such great power must not fall into ignorant hands."

"That bleedin' apprentice," I grouse under my breath, "he's been yakkin' to the villagers again." Then I look at Mark's stern face, and sigh. "They aren't magic pebbles," I say aloud. "Just ordinary stones I picked up from the ground."

A flicker of confusion crosses Mark's face, then he brightens again. "I'm here for the magic bucket!" he declares.

"It's not a magic bucket," I say wearily. "I used to keep dirty socks in it."

Mark's face is puzzled. "Then where is the magic?" he demands.

An interesting question. "It's hard to explain," I say.

My current apprentice, Autrey, attracted by the commotion, wanders over and volunteers his explanation: "It's the level of pebbles in the bucket," Autrey says. "There's a magic level of pebbles, and you have to get the level just right, or it doesn't work. If you throw in more pebbles, or take some out, the bucket won't be at the magic level anymore. Right now, the magic level is," Autrey peers into the bucket, "about one-third full."

"I see!" Mark says excitedly. From his back pocket Mark takes out his own bucket, and a heap of pebbles. Then he grabs a few handfuls of pebbles, and stuffs them into the bucket. Then Mark looks into the bucket, noting how many pebbles are there. "There we go," Mark says, "the magic level of this bucket is half full. Like that?"

"No!" Autrey says sharply. "Half full is not the magic level. The magic level is about one-third. Half full is definitely unmagic. Furthermore, you're using the wrong bucket."

Mark turns to me, puzzled. "I thought you said the bucket wasn't magic?"

"It's not," I say. A sheep passes out through the gate, and I toss another pebble into the bucket. "Besides, I'm watching the sheep. Talk to Autrey."

Mark dubiously eyes the pebble I tossed in, but decides to temporarily shelve the question. Mark turns to Autrey and draws himself up haughtily. "It's a free country," Mark says, "under the benev-

olent dictatorship of the Senate, of course. I can drop whichever pebbles I like into whatever bucket I like.”

Autrey considers this. “No you can’t,” he says finally, “there won’t be any magic.”

“Look,” says Mark patiently, “I watched you carefully. You looked in your bucket, checked the level of pebbles, and called that the magic level. I did exactly the same thing.”

“That’s not how it works,” says Autrey.

“Oh, I see,” says Mark, “It’s not the level of pebbles in *my* bucket that’s magic, it’s the level of pebbles in *your* bucket. Is that what you claim? What makes your bucket so much better than mine, huh?”

“Well,” says Autrey, “if we were to empty your bucket, and then pour all the pebbles from my bucket into your bucket, then your bucket would have the magic level. There’s also a procedure we can use to check if your bucket has the magic level, if we know that my bucket has the magic level; we call that a bucket compare operation.”

Another sheep passes, and I toss in another pebble.

“He just tossed in another pebble!” Mark says. “And I suppose you claim the new level is also magic? I could toss pebbles into your bucket until the level was the same as mine, and then our buckets would agree. You’re just comparing my bucket to your bucket to determine whether *you* think the level is ‘magic’ or not. Well, I think *your* bucket isn’t magic, because it doesn’t have the same level of pebbles as mine. So there!”

“Wait,” says Autrey, “you don’t understand -”

“By ‘magic level’, you mean simply the level of pebbles in *your* own bucket. And when I say ‘magic level’, I mean the level of pebbles in *my* bucket. Thus you look at my bucket and say it ’isn’t magic’, but the word ‘magic’ means different things to different people. You need to specify *whose* magic it is. You should say that my bucket doesn’t have ‘Autrey’s magic level’, and I say that your bucket doesn’t have ‘Mark’s magic level’. That way, the apparent contradiction goes away.”

“But -” says Autrey helplessly.

"Different people can have different buckets with different levels of pebbles, which proves this business about 'magic' is completely arbitrary and subjective."

"Mark," I say, "did anyone tell you what these pebbles *do*?"

"*Do?*?" says Mark. "I thought they were just magic."

"If the pebbles didn't do anything," says Autrey, "our ISO 9000 process efficiency auditor would eliminate the procedure from our daily work."

"What's your auditor's name?"

"Darwin," says Autrey.

"Hm," says Mark. "Charles does have a reputation as a strict auditor. So do the pebbles bless the flocks, and cause the increase of sheep?"

"No," I say. "The virtue of the pebbles is this; if we look into the bucket and see the bucket is empty of pebbles, we know the pastures are likewise empty of sheep. If we do not use the bucket, we must search and search until dark, lest one last sheep remain. Or if we stop our work early, then sometimes the next morning we find a dead sheep, for the wolves savage any sheep left outside. If we look in the bucket, we know when all the sheep are home, and we can retire without fear."

Mark considers this. "That sounds rather implausible," he says eventually. "Did you consider using divination sticks? Divination sticks are infallible, or at least, anyone who says they are fallible is burned at the stake. This is an extremely painful way to die; it follows that divination sticks are infallible."

"You're welcome to use divination sticks if you like," I say.

"Oh, good heavens, of course not," says Mark. "They work infallibly, with absolute perfection on every occasion, as befits such blessed instruments; but what if there were a dead sheep the next morning? I only use the divination sticks when there is no possibility of their being proven wrong. Otherwise I might be burned alive. So how does your magic bucket work?"

How does the bucket work...? I'd better start with the simplest possible case. "Well," I say, "suppose the pastures are empty, and the bucket isn't empty. Then we'll waste hours looking for a sheep that isn't there. And if there are sheep in the pastures, but the bucket is empty, then Autrey and I will turn in too early, and we'll find

dead sheep the next morning. So an empty bucket is magical if and only if the pastures are empty –”

“Hold on,” says Autrey. “That sounds like a vacuous tautology to me. Aren’t an empty bucket and empty pastures obviously the same thing?”

“It’s not vacuous,” I say. “Here’s an analogy: The logician Alfred Tarski once said that the assertion ‘Snow is white’ is true if and only if snow is white. If you can understand that, you should be able to see why an empty bucket is magical if and only if the pastures are empty of sheep.”

“Hold on,” says Mark. “These are *buckets*. They don’t have anything to do with *sheep*. Buckets and sheep are obviously completely different. There’s no way the sheep can ever interact with the bucket.”

“Then where do *you* think the magic comes from?” inquires Autrey.

Mark considers. “You said you could compare two buckets to check if they had the same level... I can see how buckets can interact with buckets. Maybe when you get a large collection of buckets, and they all have the same level, *that’s* what generates the magic. I’ll call that the coherentist theory of magic buckets.”

“Interesting,” says Autrey. “I know that my master is working on a system with multiple buckets – he says it might work better because of ‘redundancy’ and ‘error correction’. That sounds like coherentism to me.”

“They’re not quite the same –” I start to say.

“Let’s test the coherentism theory of magic,” says Autrey. “I can see you’ve got five more buckets in your back pocket. I’ll hand you the bucket we’re using, and then you can fill up your other buckets to the same level –”

Mark recoils in horror. “Stop! These buckets have been passed down in my family for generations, and they’ve always had the same level! If I accept your bucket, my bucket collection will become less coherent, and the magic will go away!”

“But your *current* buckets don’t have anything to do with the sheep!” protests Autrey.

Mark looks exasperated. “Look, I’ve explained before, there’s obviously no way that sheep can interact with buckets. Buckets can only interact with other buckets.”

“I toss in a pebble whenever a sheep passes,” I point out.

“When a sheep passes, you toss in a pebble?” Mark says. “What does that have to do with anything?”

“It’s an interaction between the sheep and the pebbles,” I reply.

“No, it’s an interaction between the pebbles and *you*,” Mark says. “The magic doesn’t come from the sheep, it comes from *you*. Mere sheep are obviously nonmagical. The magic has to come from *somewhere*, on the way to the bucket.”

I point at a wooden mechanism perched on the gate. “Do you see that flap of cloth hanging down from that wooden contraption? We’re still fiddling with that – it doesn’t work reliably – but when sheep pass through, they disturb the cloth. When the cloth moves aside, a pebble drops out of a reservoir and falls into the bucket. That way, Autrey and I won’t have to toss in the pebbles ourselves.”

Mark furrows his brow. “I don’t quite follow you... is the *cloth* magical?”

I shrug. “I ordered it online from a company called Natural Selections. The fabric is called Sensory Modality.” I pause, seeing the incredulous expressions of Mark and Autrey. “I admit the names are a bit New Agey. The point is that a passing sheep triggers a chain of cause and effect that ends with a pebble in the bucket. *Afterward* you can compare the bucket to other buckets, and so on.”

“I still don’t get it,” Mark says. “You can’t fit a sheep into a bucket. Only pebbles go in buckets, and it’s obvious that pebbles only interact with other pebbles.”

“The sheep interact with things that interact with pebbles...” I search for an analogy. “Suppose you look down at your shoelaces. A photon leaves the Sun; then travels down through Earth’s atmosphere; then bounces off your shoelaces; then passes through the pupil of your eye; then strikes the retina; then is absorbed by a rod or a cone. The photon’s energy makes the attached neuron fire, which causes other neurons to fire. A neural activation pattern in your visual cortex can interact with your beliefs about your shoelaces, since beliefs about shoelaces also exist in neural sub-

strate. If you can understand that, you should be able to see how a passing sheep causes a pebble to enter the bucket.”

“At exactly *which* point in the process does the pebble become magic?” says Mark.

“It... um...” Now *I’m* starting to get confused. I shake my head to clear away cobwebs. This all seemed simple enough when I woke up this morning, and the pebble-and-bucket system hasn’t gotten any more complicated since then. “This is a lot easier to understand if you remember that the *point* of the system is to keep track of sheep.”

Mark sighs sadly. “Never mind... it’s obvious you don’t know. Maybe all pebbles are magical to start with, even before they enter the bucket. We could call that position panpebblism.”

“Ha!” Autrey says, scorn rich in his voice. “Mere wishful thinking! Not all pebbles are created equal. The pebbles in *your* bucket are *not* magical. They’re only lumps of stone!”

Mark’s face turns stern. “Now,” he cries, “now you see the danger of the road you walk! Once you say that some people’s pebbles are magical and some are not, your pride will consume you! You will think yourself superior to all others, and so fall! Many throughout history have tortured and murdered because they thought their own pebbles supreme!” A tinge of condescension enters Mark’s voice. “Worshipping a level of pebbles as ‘magical’ implies that there’s an absolute pebble level in a Supreme Bucket. Nobody believes in a Supreme Bucket these days.”

“One,” I say. “Sheep are not absolute pebbles. Two, I don’t think my bucket actually contains the sheep. Three, I don’t worship my bucket level as perfect – I adjust it sometimes – and I do that because I care about the sheep.”

“Besides,” says Autrey, “someone who believes that possessing absolute pebbles *would* license torture and murder, is making a mistake that has nothing to do with buckets. You’re solving the wrong problem.”

Mark calms himself down. “I suppose I can’t expect any better from mere shepherds. You probably believe that snow is white, don’t you.”

“Um... yes?” says Autrey.

“It doesn’t bother you that *Joseph Stalin* believed that snow is white?”

“Um... no?” says Autrey.

Mark gazes incredulously at Autrey, and finally shrugs. “Let’s suppose, purely for the sake of argument, that your pebbles are magical and mine aren’t. Can you tell me what the difference is?”

“My pebbles *represent* the sheep!” Autrey says triumphantly. “*Your* pebbles don’t have the representativeness property, so they won’t work. They are empty of meaning. Just look at them. There’s no aura of semantic content; they are merely pebbles. You need a bucket with special causal powers.”

“Ah!” Mark says. “Special causal powers, instead of magic.”

“Exactly,” says Autrey. “I’m not superstitious. Postulating magic, in this day and age, would be unacceptable to the international shepherding community. We have found that postulating magic simply doesn’t work as an explanation for shepherding phenomena. So when I see something I don’t understand, and I want to explain it using a model with no internal detail that makes no predictions even in retrospect, I postulate special causal powers. If that doesn’t work, I’ll move on to calling it an emergent phenomenon.”

“What kind of special powers does the bucket have?” asks Mark.

“Hm,” says Autrey. “Maybe this bucket is imbued with an *aboutness* relation to the pastures. That would explain why it worked – when the bucket is empty, it *means* the pastures are empty.”

“Where did you find this bucket?” says Mark. “And how did you realize it had an about-ness relation to the pastures?”

“It’s an *ordinary bucket*,” I say. “I used to climb trees with it... I don’t think this question *needs* to be difficult.”

“I’m talking to Autrey,” says Mark.

“You have to bind the bucket to the pastures, and the pebbles to the sheep, using a magical ritual – pardon me, an emergent process with special causal powers – that my master discovered,” Autrey explains.

Autrey then attempts to describe the ritual, with Mark nodding along in sage comprehension.

"You have to throw in a pebble *every* time a sheep leaves through the gate?" says Mark. "Take out a pebble *every* time a sheep returns?"

Autrey nods. "Yeah."

"That must be really hard," Mark says sympathetically.

Autrey brightens, soaking up Mark's sympathy like rain. "Exactly!" says Autrey. "It's *extremely* hard on your emotions. When the bucket has held its level for a while, you... tend to get attached to that level."

A sheep passes then, leaving through the gate. Autrey sees; he stoops, picks up a pebble, holds it aloft in the air. "Behold!" Autrey proclaims. "A sheep has passed! I must now toss a pebble into this bucket, my dear bucket, and destroy that fond level which has held for so long –" Another sheep passes. Autrey, caught up in his drama, misses it; so I plunk a pebble into the bucket. Autrey is still speaking: " – for that is the supreme test of the shepherd, to throw in the pebble, be it ever so agonizing, be the old level ever so precious. Indeed, only the best of shepherds can meet a requirement so stern –"

"Autrey," I say, "if you want to be a great shepherd someday, learn to shut up and throw in the pebble. No fuss. No drama. Just do it."

"And this ritual," says Mark, "it binds the pebbles to the sheep by the magical laws of Sympathy and Contagion, like a voodoo doll."

Autrey winces and looks around. "Please! Don't call it Sympathy and Contagion. We shepherds are an anti-superstitious folk. Use the word 'intentionality', or something like that."

"Can I look at a pebble?" says Mark.

"Sure," I say. I take one of the pebbles out of the bucket, and toss it to Mark. Then I reach to the ground, pick up another pebble, and drop it into the bucket.

Autrey looks at me, puzzled. "Didn't you just mess it up?"

I shrug. "I don't think so. We'll know I messed it up if there's a dead sheep next morning, or if we search for a few hours and don't find any sheep."

"But –" Autrey says.

"I taught you everything *you* know, but I haven't taught you everything *I* know," I say.

Mark is examining the pebble, staring at it intently. He holds his hand over the pebble and mutters a few words, then shakes his head. "I don't sense any magical power," he says. "Pardon me. I don't sense any intentionality."

"A pebble only has intentionality if it's inside a ma- an emergent bucket," says Autrey. "Otherwise it's just a mere pebble."

"Not a problem," I say. I take a pebble out of the bucket, and toss it away. Then I walk over to where Mark stands, tap his hand holding a pebble, and say: "I declare this hand to be part of the mag- ic bucket!" Then I resume my post at the gates.

Autrey laughs. "Now you're just being gratuitously evil."

I nod, for this is indeed the case.

"Is that really going to work, though?" says Autrey.

I nod again, hoping that I'm right. I've done this before with two buckets, and in principle, there should be no difference between Mark's hand and a bucket. Even if Mark's hand is imbued with the *elan vital* that distinguishes live matter from dead matter, the trick should work as well as if Mark were a marble statue.

Mark is looking at his hand, a bit unnerved. "So... the pebble has intentionality again, now?"

"Yep," I say. "Don't add any more pebbles to your hand, or throw away the one you have, or you'll break the ritual."

Mark nods solemnly. Then he resumes inspecting the pebble. "I understand now how your flocks grew so great," Mark says. "With the power of this bucket, you could keep in tossing pebbles, and the sheep would keep returning from the fields. You could start with just a few sheep, let them leave, then fill the bucket to the brim before they returned. And if tending so many sheep grew tedious, you could let them all leave, then empty almost all the pebbles from the bucket, so that only a few returned... increasing the flocks again when it came time for shearing... dear heavens, man! Do you realize the sheer *power* of this ritual you've discovered? I can only imagine the implications; humankind might leap ahead a decade – no, a century!"

"It doesn't work that way," I say. "If you add a pebble when a sheep hasn't left, or remove a pebble when a sheep hasn't come in,

that breaks the ritual. The power does not linger in the pebbles, but vanishes all at once, like a soap bubble popping.”

Mark’s face is terribly disappointed. “Are you sure?”

I nod. “I tried that and it didn’t work.”

Mark sighs heavily. “And this... *math*... seemed so powerful and useful until then... Oh, well. So much for human progress.”

“Mark, it was a *brilliant* idea,” Autrey says encouragingly. “The notion didn’t occur to me, and yet it’s so obvious... it would save an *enormous* amount of effort... there *must* be a way to salvage your plan! We could try different buckets, looking for one that would keep the magical pow- the intentionality in the pebbles, even without the ritual. Or try other pebbles. Maybe our pebbles just have the wrong properties to have *inherent* intentionality. What if we tried it using stones carved to resemble tiny sheep? Or just write ‘sheep’ on the pebbles; that might be enough.”

“Not going to work,” I predict dryly.

Autrey continues. “Maybe we need organic pebbles, instead of silicon pebbles... or maybe we need to use expensive gemstones. The price of gemstones doubles every eighteen months, so you could buy a handful of cheap gemstones now, and wait, and in twenty years they’d be really expensive.”

“You tried adding pebbles to create more sheep, and it didn’t work?” Mark asks me. “What exactly did you do?”

“I took a handful of dollar bills. Then I hid the dollar bills under a fold of my blanket, one by one; each time I hid another bill, I took another paperclip from a box, making a small heap. I was careful not to keep track in my head, so that all I knew was that there were ‘many’ dollar bills, and ‘many’ paperclips. Then when all the bills were hidden under my blanket, I added a single additional paperclip to the heap, the equivalent of tossing an extra pebble into the bucket. Then I started taking dollar bills from under the fold, and putting the paperclips back into the box. When I finished, a single paperclip was left over.”

“What does that result mean?” asks Autrey.

“It means the trick didn’t work. Once I broke ritual by that single misstep, the power did not linger, but vanished instantly; the heap of paperclips and the pile of dollar bills no longer went empty at the same time.”

“You *actually* tried this?” asks Mark.

“Yes,” I say, “I actually performed the experiment, to verify that the outcome matched my theoretical prediction. I have a sentimental fondness for the scientific method, even when it seems absurd. Besides, what if I’d been wrong?”

“If it *had* worked,” says Mark, “you would have been guilty of counterfeiting! Imagine if everyone did that; the economy would collapse! Everyone would have billions of dollars of currency, yet there would be nothing for money to buy!”

“Not at all,” I reply. “By that same logic whereby adding another paperclip to the heap creates another dollar bill, creating another dollar bill would create an additional dollar’s worth of goods and services.”

Mark shakes his head. “Counterfeiting is still a crime... You should not have tried.”

“I was *reasonably* confident I would fail.”

“Aha!” says Mark. “You *expected* to fail! You didn’t *believe* you could do it!”

“Indeed,” I admit. “You have guessed my expectations with stunning accuracy.”

“Well, that’s the problem,” Mark says briskly. “Magic is fueled by belief and willpower. If you don’t believe you can do it, you can’t. You need to change your belief about the experimental result; that will change the result itself.”

“Funny,” I say nostalgically, “that’s what Autrey said when I told him about the pebble-and-bucket method. That it was too ridiculous for him to believe, so it wouldn’t work for him.”

“How did you persuade him?” inquires Mark.

“I told him to shut up and follow instructions,” I say, “and when the method worked, Autrey started believing in it.”

Mark frowns, puzzled. “That makes no sense. It doesn’t resolve the essential chicken-and-egg dilemma.”

“Sure it does. The bucket method works whether or not you believe in it.”

“That’s *absurd!*” sputters Mark. “I don’t believe in magic that works whether or not you believe in it!”

“I said that too,” chimes in Autrey. “Apparently I was wrong.”

Mark screws up his face in concentration. “But... if you didn’t believe in magic that works whether or not you believe in it, then why did the bucket method work when you didn’t believe in it? Did you believe in magic that works whether or not you believe in it whether or not you believe in magic that works whether or not you believe in it?”

“I don’t... *think* so...” says Autrey doubtfully.

“Then if you didn’t believe in magic that works whether or not you... hold on a second, I need to work this out on paper and pencil -” Mark scribbles frantically, looks skeptically at the result, turns the piece of paper upside down, then gives up. “Never mind,” says Mark. “Magic is difficult enough for me to comprehend; metamagic is out of my depth.”

“Mark, I don’t think you understand the art of bucketcraft,” I say. “It’s not about using pebbles to control sheep. It’s about making sheep control pebbles. In this art, it is not necessary to begin by believing the art will work. Rather, first the art works, then one comes to believe that it works.”

“Or so you believe,” says Mark.

“So I believe,” I reply, “*because* it happens to be a fact. The correspondence between reality and my beliefs comes from reality controlling my beliefs, not the other way around.”

Another sheep passes, causing me to toss in another pebble.

“Ah! Now we come to the root of the problem,” says Mark. “What’s this so-called ‘reality’ business? I understand what it means for a hypothesis to be elegant, or falsifiable, or compatible with the evidence. It sounds to me like calling a belief ‘true’ or ‘real’ or ‘actual’ is merely the difference between saying you believe something, and saying you really really believe something.”

I pause. “Well...” I say slowly. “Frankly, I’m not entirely sure myself where this ‘reality’ business comes from. I can’t create my own reality in the lab, so I must not understand it yet. But occasionally I believe strongly that something is going to happen, and then something else happens instead. I need a name for whatever-it-is that determines my experimental results, so I call it ‘reality’. This ‘reality’ is somehow separate from even my very best hypotheses. Even when I have a simple hypothesis, strongly supported by all the evidence I know, sometimes I’m still surprised. So I need differ-

ent names for the thingies that determine my predictions and the thingy that determines my experimental results. I call the former thingies ‘belief’, and the latter thingy ‘reality’.”

Mark snorts. “I don’t even know why I bother listening to this obvious nonsense. Whatever you say about this so-called ‘reality’, it is merely another belief. Even your belief that reality precedes your beliefs is a belief. It follows, as a logical inevitability, that reality does not exist; only beliefs exist.”

“Hold on,” says Autrey, “could you repeat that last part? You lost me with that sharp swerve there in the middle.”

“No matter what you say about reality, it’s just another belief,” explains Mark. “It follows with crushing necessity that there is no reality, only beliefs.”

“I see,” I say. “The same way that no matter what you eat, you need to eat it with your mouth. It follows that there is no food, only mouths.”

“Precisely,” says Mark. “Everything that you eat has to be in your mouth. How can there be food that exists outside your mouth? The thought is nonsense, proving that ‘food’ is an incoherent notion. That’s why we’re all starving to death; there’s no food.”

Autrey looks down at his stomach. “But I’m *not* starving to death.”

“*Aha!*” shouts Mark triumphantly. “And how did you utter that very objection? With your *mouth*, my friend! With your *mouth!* What better demonstration could you ask that there is no food?”

“*What’s this about starvation?*” demands a harsh, rasping voice from directly behind us. Autrey and I stay calm, having gone through this before. Mark leaps a foot in the air, startled almost out of his wits.

Inspector Darwin smiles tightly, pleased at achieving surprise, and makes a small tick on his clipboard.

“Just a metaphor!” Mark says quickly. “You don’t need to take away my mouth, or anything like that –”

“*Why do you need a mouth if there is no food?*” demands Darwin angrily. “*Never mind.* I have no *time* for this *foolishness*. I am here to inspect the *sheep*.”

“Flocks thriving, sir,” I say. “No dead sheep since January.”

"Excellent. I award you 0.12 units of *fitness*. Now what is this *person* doing here? Is he a necessary part of the *operations*?"

"As far as I can see, he would be of more use to the human species if hung off a hot-air balloon as ballast," I say.

"Ouch," says Autrey mildly.

"I do not *care* about the *human species*. Let him speak for *himself*."

Mark draws himself up haughtily. "This mere *shepherd*," he says, gesturing at me, "has claimed that there is such a thing as reality. This offends me, for I know with deep and abiding certainty that there is no truth. The concept of 'truth' is merely a stratagem for people to impose their own beliefs on others. Every culture has a different 'truth', and no culture's 'truth' is superior to any other. This that I have said holds at all times in all places, and I insist that you agree."

"Hold on a second," says Autrey. "If nothing is true, why should I believe you when you say that nothing is true?"

"I didn't say that nothing is true –" says Mark.

"Yes, you did," interjects Autrey, "I heard you."

"– I said that 'truth' is an excuse used by some cultures to enforce their beliefs on others. So when you say something is 'true', you mean only that it would be advantageous to your own social group to have it believed."

"And this that you have said," I say, "is it true?"

"Absolutely, positively true!" says Mark emphatically. "People create their own realities."

"Hold on," says Autrey, sounding puzzled again, "saying that people create their own realities is, logically, a completely separate issue from saying that there is no truth, a state of affairs I cannot even imagine coherently, perhaps because you still have not explained how exactly it is supposed to work –"

"There you go again," says Mark exasperatedly, "trying to apply your Western concepts of logic, rationality, reason, coherence, and self-consistency."

"Great," mutters Autrey, "now I need to add a *third* subject heading, to keep track of this entirely separate and distinct claim –"

"It's not separate," says Mark. "Look, you're taking the wrong attitude by treating my statements as hypotheses, and carefully

deriving their consequences. You need to think of them as fully general excuses, which I apply when anyone says something I don't like. It's not so much a model of how the universe works, as a "Get Out of Jail Free" card. The *key* is to apply the excuse *selectively*. When I say that there is no such thing as truth, that applies only to *your* claim that the magic bucket works whether or not I believe in it. It does *not* apply to *my* claim that there is no such thing as truth."

"Um... why not?" inquires Autrey.

Mark heaves a patient sigh. "Autrey, do you think you're the first person to think of that question? To ask us how our own beliefs can be meaningful if all beliefs are meaningless? That's the same thing many students say when they encounter this philosophy, which, I'll have you know, has many adherents and an extensive literature."

"So what's the answer?" says Autrey.

"We named it the 'reflexivity problem,'" explains Mark.

"But what's the *answer*?" persists Autrey.

Mark smiles condescendingly. "Believe me, Autrey, you're not the first person to think of such a simple question. There's no point in presenting it to us as a triumphant refutation."

"But what's the *actual answer*?"

"Now, I'd like to move on to the issue of how logic kills cute baby seals -"

"*You* are wasting *time*," snaps Inspector Darwin.

"Not to mention, losing track of sheep," I say, tossing in another pebble.

Inspector Darwin looks at the two arguers, both apparently unwilling to give up their positions. "Listen," Darwin says, more kindly now, "I have a simple notion for resolving your dispute. *You* say," says Darwin, pointing to Mark, "that people's beliefs alter their personal realities. And *you* fervently believe," his finger swivels to point at Autrey, "that Mark's beliefs *can't* alter reality. So let Mark believe really hard that he can fly, and then step off a cliff. Mark shall see himself fly away like a bird, and Autrey shall see him plummet down and go splat, and you shall both be happy."

We all pause, considering this.

"It *sounds* reasonable..." Mark says finally.

"There's a cliff right there," observes Inspector Darwin.

Autrey is wearing a look of intense concentration. Finally he shouts: "Wait! If that were true, we would all have long since departed into our own private universes, in which case the other people here are only figments of your imagination – there's no point in trying to prove anything to us –"

A long dwindling scream comes from the nearby cliff, followed by a dull and lonely splat. Inspector Darwin flips his clipboard to the page that shows the current gene pool and pencils in a slightly lower frequency for Mark's alleles.

Autrey looks slightly sick. "Was that really necessary?"

"*Necessary?*" says Inspector Darwin, sounding puzzled. "It just *happened...* I don't quite understand your question."

Autrey and I turn back to our bucket. It's time to bring in the sheep. You wouldn't want to forget about that part. Otherwise what would be the point?

## 2. What Do We Mean By “Rationality”?<sup>1</sup>

We mean:

1. **Epistemic rationality:** believing, and updating on evidence, so as to systematically improve the correspondence between your map and the territory<sup>2</sup>. 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<sup>3</sup>, 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  $P(A) \geq P(A \& B)$ , the judgment  $P(\text{"Bill plays jazz"}) < P(\text{"Bill plays jazz"} \& \text{"Bill is accountant"})$  is labeled incorrect.

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:

---

1. [http://lesswrong.com/lw/31/what\\_do\\_we\\_mean\\_by\\_rationality/](http://lesswrong.com/lw/31/what_do_we_mean_by_rationality/)

2. Page 5, ‘The Simple Truth’.

3. [http://lesswrong.com/lw/ji/conjunction\\_fallacy/](http://lesswrong.com/lw/ji/conjunction_fallacy/)

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<sup>4</sup> in which the number of observers is uncertain. The exact rules of decision theory are called into question by e.g. Newcomblike problems<sup>5</sup> 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 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<sup>6</sup> and here<sup>7</sup>.)

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<sup>8</sup>. Here again, trying to define “rationality” a particular way doesn’t support an answer, merely presume it.

From the Twelve Virtues of Rationality<sup>9</sup>:

---

4. <http://www.anthropic-principle.com/primer.html>

5. [http://lesswrong.com/lw/nc/newcombs\\_problem\\_and\\_regret\\_of\\_rationality/](http://lesswrong.com/lw/nc/newcombs_problem_and_regret_of_rationality/)

6. Page 193, ‘The Parable of Hemlock’.

7. Page 263, ‘Arguing “By Definition”’.

8. Page 625, ‘Feeling Rational’.

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 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<sup>10</sup>, not even if that word is “rationality”. The point of attaching sequences of letters to particular concepts is to let two people *communicate*<sup>11</sup> - 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 means*

9. <http://yudkowsky.net/virtues/>

10. Page 228, ‘Disputing Definitions’.

11. Page 238, ‘The Argument from Common Usage’.

*ing* 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.

In this case - or in any other case where controversy threatens - you should substitute more specific language<sup>12</sup>: “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<sup>13</sup>, 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. Page 245, ‘Taboo Your Words’.

13. [http://lesswrong.com/lw/nc/newcombs\\_problem\\_and\\_regret\\_of\\_rationality/](http://lesswrong.com/lw/nc/newcombs_problem_and_regret_of_rationality/)

### **3. An Intuitive Explanation of Bayes' Theorem<sup>1</sup>**

*Bayes' Theorem  
for the curious and bewildered;  
an excruciatingly gentle introduction.*

---

Your friends and colleagues are talking about something called “Bayes’ Theorem” or “Bayes’ Rule”, or something called Bayesian reasoning. They sound really enthusiastic about it, too, so you google and find a webpage about Bayes’ Theorem and...

It’s this equation. That’s all. Just one equation. The page you found gives a definition of it, but it doesn’t say what it is, or why it’s useful, or why your friends would be interested in it. It looks like this random statistics thing.

So you came here. Maybe you don’t understand what the equation says. Maybe you understand it in theory, but every time you try to apply it in practice you get mixed up trying to remember the difference between  $p(a|x)$  and  $p(x|a)$ , and whether  $p(a) * p(x|a)$  belongs in the numerator or the denominator. Maybe you see the theorem, and you understand the theorem, and you can use the theorem, but you can’t understand why your friends and/or research colleagues seem to think it’s the secret of the universe. Maybe your friends are all wearing Bayes’ Theorem T-shirts, and you’re feeling left out. Maybe you’re a girl looking for a boyfriend, but the boy you’re interested in refuses to date anyone who “isn’t Bayesian”. What matters is that Bayes is cool, and if you don’t know Bayes, you aren’t cool.

Why does a mathematical concept generate this strange enthusiasm in its students? What is the so-called Bayesian Revolution now sweeping through the sciences, which claims to subsume even the experimental method itself as a special case? What is the secret that the adherents of Bayes know? What is the light that they have seen?

---

<sup>1</sup>. <http://yudkowsky.net/rational/bayes>

Soon you will know. Soon you will be one of us.

While there are a few existing online explanations of Bayes' Theorem, my experience with trying to introduce people to Bayesian reasoning is that the existing online explanations are too abstract. Bayesian reasoning is very *counterintuitive*. People do not employ Bayesian reasoning intuitively, find it very difficult to learn Bayesian reasoning when tutored, and rapidly forget Bayesian methods once the tutoring is over. This holds equally true for novice students and highly trained professionals in a field. Bayesian reasoning is apparently one of those things which, like quantum mechanics or the Wason Selection Test, is inherently difficult for humans to grasp with our built-in mental faculties.

Or so they claim. Here you will find an attempt to offer an *intuitive* explanation of Bayesian reasoning - an excruciatingly gentle introduction that invokes all the human ways of grasping numbers, from natural frequencies to spatial visualization. The intent is to convey, not abstract rules for manipulating numbers, but what the numbers mean, and why the rules are what they are (and cannot possibly be anything else). When you are finished reading this page, you will see Bayesian problems in your dreams.

And let's begin.

---

---

Here's a story problem about a situation that doctors often encounter:

1% of women at age forty who participate in routine screening have breast cancer. 80% of women with breast cancer will get positive mammographies. 9.6% of women without breast cancer will also get positive mammographies. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?

What do you think the answer is? If you haven't encountered this kind of problem before, please take a moment to come up with

your own answer before continuing.

---

Next, suppose I told you that most doctors get the same wrong answer on this problem - usually, only around 15% of doctors get it right. ("Really? 15%? Is that a real number, or an urban legend based on an Internet poll?" It's a real number. See Casscells, Schoenberger, and Grayboys 1978; Eddy 1982; Gigerenzer and Hoffrage 1995; and many other studies. It's a surprising result which is easy to replicate, so it's been extensively replicated.)

Do you want to think about your answer again? Here's a Javascript calculator if you need one. This calculator has the usual precedence rules; multiplication before addition and so on. If you're not sure, I suggest using parentheses.

---

On the story problem above, most doctors estimate the probability to be between 70% and 80%, which is wildly incorrect.

Here's an alternate version of the problem on which doctors fare somewhat better:

10 out of 1000 women at age forty who participate in routine screening have breast cancer. 800 out of 1000 women with breast cancer will get positive mammographies. 96 out of 1000 women without breast cancer will also get positive mammographies. If 1000 women in this age group undergo a routine screening, about what fraction of women with positive mammographies will actually have breast cancer?

---

And finally, here's the problem on which doctors fare best of all, with 46% - nearly half - arriving at the correct answer:

100 out of 10,000 women at age forty who participate in routine screening have breast cancer. 80 of every 100

women with breast cancer will get a positive mammography. 950 out of 9,900 women without breast cancer will also get a positive mammography. If 10,000 women in this age group undergo a routine screening, about what fraction of women with positive mammographies will actually have breast cancer?

---

The correct answer is 7.8%, obtained as follows: Out of 10,000 women, 100 have breast cancer; 80 of those 100 have positive mammographies. From the same 10,000 women, 9,900 will not have breast cancer and of those 9,900 women, 950 will also get positive mammographies. This makes the total number of women with positive mammographies  $950 + 80$  or 1,030. Of those 1,030 women with positive mammographies, 80 will have cancer. Expressed as a proportion, this is  $80/1,030$  or 0.07767 or 7.8%.

To put it another way, before the mammography screening, the 10,000 women can be divided into two groups:

- Group 1: 100 women *with* breast cancer.
- Group 2: 9,900 women *without* breast cancer.

Summing these two groups gives a total of 10,000 patients, confirming that none have been lost in the math. After the mammography, the women can be divided into four groups:

- Group A: 80 women *with* breast cancer, and a *positive* mammography.
- Group B: 20 women *with* breast cancer, and a *negative* mammography.
- Group C: 950 women *without* breast cancer, and a *positive* mammography.
- Group D: 8,950 women *without* breast cancer, and a *negative* mammography.

As you can check, the sum of all four groups is still 10,000. The sum of groups A and B, the groups with breast cancer, corresponds to group 1; and the sum of groups C and D, the groups without breast cancer, corresponds to group 2; so administering a mammography does not actually *change* the number of women with breast cancer. The proportion of the cancer patients (A + B) within the complete set of patients (A + B + C + D) is the same as

the 1% prior chance that a woman has cancer:  $(80 + 20) / (80 + 20 + 950 + 8950) = 100 / 10000 = 1\%$ .

The proportion of cancer patients with positive results, within the group of *all* patients with positive results, is the proportion of (A) within (A + C):  $80 / (80 + 950) = 80 / 1030 = 7.8\%$ . If you administer a mammography to 10,000 patients, then out of the 1030 with positive mammographies, 80 of those positive-mammography patients will have cancer. This is the correct answer, the answer a doctor should give a positive-mammography patient if she asks about the chance she has breast cancer; if thirteen patients ask this question, roughly 1 out of those 13 will have cancer.

---

The most common mistake is to ignore the original fraction of women with breast cancer, and the fraction of women without breast cancer who receive false positives, and focus only on the fraction of women with breast cancer who get positive results. For example, the vast majority of doctors in these studies seem to have thought that if around 80% of women with breast cancer have positive mammographies, then the probability of a women with a positive mammography having breast cancer must be around 80%.

Figuring out the final answer always requires *all three* pieces of information - the percentage of women with breast cancer, the percentage of women without breast cancer who receive false positives, and the percentage of women with breast cancer who receive (correct) positives.

To see that the final answer always depends on the original fraction of women with breast cancer, consider an alternate universe in which only one woman out of a million has breast cancer. Even if mammography in this world detects breast cancer in 8 out of 10 cases, while returning a false positive on a woman without breast cancer in only 1 out of 10 cases, there will still be a hundred thousand false positives for every real case of cancer detected. The original probability that a woman has cancer is so extremely low that, although a positive result on the

mammography does *increase* the estimated probability, the probability isn't increased to certainty or even "a noticeable chance"; the probability goes from 1:1,000,000 to 1:100,000.

Similarly, in an alternate universe where only one out of a million women does *not* have breast cancer, a positive result on the patient's mammography obviously doesn't mean that she has an 80% chance of having breast cancer! If this were the case her estimated probability of having cancer would have been revised drastically *downward* after she got a *positive* result on her mammography - an 80% chance of having cancer is a lot less than 99.9999%! If you administer mammographies to ten million women in this world, around eight million women with breast cancer will get correct positive results, while one woman without breast cancer will get false positive results. Thus, if you got a positive mammography in this alternate universe, your chance of having cancer would go from 99.9999% up to 99.999987%. That is, your chance of being healthy would go from 1:1,000,000 down to 1:8,000,000.

These two extreme examples help demonstrate that the mammography result doesn't *replace* your old information about the patient's chance of having cancer; the mammography *slides* the estimated probability in the direction of the result. A positive result slides the original probability upward; a negative result slides the probability downward. For example, in the original problem where 1% of the women have cancer, 80% of women with cancer get positive mammographies, and 9.6% of women without cancer get positive mammographies, a positive result on the mammography *slides* the 1% chance upward to 7.8%.

Most people encountering problems of this type for the first time carry out the mental operation of *replacing* the original 1% probability with the 80% probability that a woman with cancer gets a positive mammography. It may seem like a good idea, but it just doesn't work. "The probability that a woman with a positive mammography has breast cancer" is not at all the same thing as "the probability that a woman with breast cancer has a positive mammography"; they are as unlike as apples and cheese. Finding

the final answer, “the probability that a woman with a positive mammography has breast cancer”, uses all three pieces of problem information - “the prior probability that a woman has breast cancer”, “the probability that a woman with breast cancer gets a positive mammography”, and “the probability that a woman without breast cancer gets a positive mammography”.

---

**Fun Fact!****Q. What is the Bayesian Conspiracy?**

A. The Bayesian Conspiracy is a multinational, interdisciplinary, and shadowy group of scientists that controls publication, grants, tenure, and the illicit traffic in grad students. The best way to be accepted into the Bayesian Conspiracy is to join the Campus Crusade for Bayes in high school or college, and gradually work your way up to the inner circles. It is rumored that at the upper levels of the Bayesian Conspiracy exist nine silent figures known only as the Bayes Council.

---

To see that the final answer always depends on the chance that a woman *without* breast cancer gets a positive mammography, consider an alternate test, mammography+. Like the original test, mammography+ returns positive for 80% of women with breast cancer. However, mammography+ returns a positive result for only one out of a million women without breast cancer - mammography+ has the same rate of false negatives, but a vastly lower rate of false positives. Suppose a patient receives a positive mammography+. What is the chance that this patient has breast cancer? Under the new test, it is a virtual certainty - 99.988%, i.e., a 1 in 8082 chance of being healthy.

Remember, at this point, that neither mammography nor mammography+ actually *change* the number of women who have breast cancer. It may seem like “There is a virtual certainty you

have breast cancer" is a terrible thing to say, causing much distress and despair; that the more hopeful verdict of the previous mammography test - a 7.8% chance of having breast cancer - was much to be preferred. This comes under the heading of "Don't shoot the messenger". The number of women who really do have cancer stays exactly the same between the two cases. Only the accuracy with which we *detect* cancer changes. Under the previous mammography test, 80 women with cancer (who *already* had cancer, before the mammography) are first told that they have a 7.8% chance of having cancer, creating X amount of uncertainty and fear, after which more detailed tests will inform them that they definitely do have breast cancer. The old mammography test also involves informing 950 women *without* breast cancer that they have a 7.8% chance of having cancer, thus creating twelve times as much additional fear and uncertainty. The new test, mammography+, does *not* give 950 women false positives, and the 80 women with cancer are told the same facts they would have learned eventually, only earlier and without an intervening period of uncertainty. Mammography+ is thus a better test in terms of its total emotional impact on patients, as well as being more accurate. Regardless of its emotional impact, it remains a fact that a patient with positive mammography+ has a 99.988% chance of having breast cancer.

Of course, that mammography+ does *not* give 950 healthy women false positives means that all 80 of the patients with positive mammography+ will be patients with breast cancer. Thus, if you have a positive mammography+, your chance of having cancer is a virtual certainty. It is *because* mammography+ does not generate as many false positives (and needless emotional stress), that the (much smaller) group of patients who *do* get positive results will be composed almost entirely of genuine cancer patients (who have bad news coming to them regardless of when it arrives).

---

Similarly, let's suppose that we have a *less* discriminating test, mammography\*, that still has a 20% rate of false negatives, as in the original case. However, mammography\* has an 80% rate of false positives. In other words, a patient *without* breast cancer has

an 80% chance of getting a false positive result on her mammography\* test. If we suppose the same 1% prior probability that a patient presenting herself for screening has breast cancer, what is the chance that a patient with positive mammography\* has cancer?

- Group 1: 100 patients with breast cancer.
- Group 2: 9,900 patients without breast cancer.

After mammography\* screening:

- Group A: 80 patients with breast cancer and a “positive” mammography\*.
- Group B: 20 patients with breast cancer and a “negative” mammography\*.
- Group C: 7920 patients without breast cancer and a “positive” mammography\*.
- Group D: 1980 patients without breast cancer and a “negative” mammography\*.

The result works out to 80 / 8,000, or 0.01. This is exactly the same as the 1% prior probability that a patient has breast cancer! A “positive” result on mammography\* doesn’t change the probability that a woman has breast cancer at all. You can similarly verify that a “negative” mammography\* also counts for nothing. And in fact it *must* be this way, because if mammography\* has an 80% hit rate for patients with breast cancer, and also an 80% rate of false positives for patients without breast cancer, then mammography\* is completely *uncorrelated* with breast cancer. There’s no reason to call one result “positive” and one result “negative”; in fact, there’s no reason to call the test a “mammography”. You can throw away your expensive mammography\* equipment and replace it with a random number generator that outputs a red light 80% of the time and a green light 20% of the time; the results will be the same. Furthermore, there’s no reason to call the red light a “positive” result or the green light a “negative” result. You could have a green light 80% of the time and a red light 20% of the time, or a blue light 80% of the time and a purple light 20% of the time, and it would all have the same bearing on whether the patient has breast cancer: i.e., no bearing whatsoever.

We can show algebraically that this *must* hold for any case where

the chance of a true positive and the chance of a false positive are the same, i.e:

- Group 1: 100 patients with breast cancer.
- Group 2: 9,900 patients without breast cancer.

Now consider a test where the probability of a true positive and the probability of a false positive are the same number M (in the example above, M=80% or M = 0.8):

- Group A:  $100 \cdot M$  patients with breast cancer and a “positive” result.
- Group B:  $100 \cdot (1 - M)$  patients with breast cancer and a “negative” result.
- Group C:  $9,900 \cdot M$  patients without breast cancer and a “positive” result.
- Group D:  $9,900 \cdot (1 - M)$  patients without breast cancer and a “negative” result.

The proportion of patients with breast cancer, within the group of patients with a “positive” result, then equals  $100 \cdot M / (100 \cdot M + 9900 \cdot M) = 100 / (100 + 9900) = 1\%$ . This holds true regardless of whether M is 80%, 30%, 50%, or 100%. If we have a mammography\* test that returns “positive” results for 90% of patients with breast cancer and returns “positive” results for 90% of patients without breast cancer, the proportion of “positive”-testing patients who have breast cancer will still equal the original proportion of patients with breast cancer, i.e., 1%.

You can run through the same algebra, replacing the prior proportion of patients with breast cancer with an arbitrary percentage P:

- Group 1: Within some number of patients, a fraction P have breast cancer.
- Group 2: Within some number of patients, a fraction  $(1 - P)$  do not have breast cancer.

After a “cancer test” that returns “positive” for a fraction M of patients with breast cancer, and also returns “positive” for the same fraction M of patients *without* cancer:

- Group A:  $P \cdot M$  patients have breast cancer and a “positive” result.
- Group B:  $P \cdot (1 - M)$  patients have breast cancer and a “negative” result.

- Group C:  $(1 - P)^*M$  patients have no breast cancer and a “positive” result.
- Group D:  $(1 - P)^*(1 - M)$  patients have no breast cancer and a “negative” result.

The chance that a patient with a “positive” result has breast cancer is then the proportion of group A within the combined group A + C, or  $P^*M / [P^*M + (1 - P)^*M]$ , which, cancelling the common factor M from the numerator and denominator, is  $P / [P + (1 - P)]$  or  $P / 1$  or just P. If the rate of false positives is the same as the rate of true positives, you always have the same probability after the test as when you started.

Which is common sense. Take, for example, the “test” of flipping a coin; if the coin comes up heads, does it tell you anything about whether a patient has breast cancer? No; the coin has a 50% chance of coming up heads if the patient has breast cancer, and also a 50% chance of coming up heads if the patient does not have breast cancer. Therefore there is no reason to call either heads or tails a “positive” result. It’s not the probability being “50/50” that makes the coin a bad test; it’s that the two probabilities, for “cancer patient turns up heads” and “healthy patient turns up heads”, are the same. If the coin was slightly biased, so that it had a 60% chance of coming up heads, it still wouldn’t be a cancer test - what makes a coin a poor test is not that it has a 50/50 chance of coming up heads if the patient has cancer, but that it also has a 50/50 chance of coming up heads if the patient does not have cancer. You can even use a test that comes up “positive” for cancer patients 100% of the time, and still not learn anything. An example of such a test is “Add 2 + 2 and see if the answer is 4.” This test returns positive 100% of the time for patients with breast cancer. It also returns positive 100% of the time for patients without breast cancer. So you learn nothing.

The original proportion of patients with breast cancer is known as the *prior probability*. The chance that a patient with breast cancer gets a positive mammography, and the chance that a patient without breast cancer gets a positive mammography, are known as the two *conditional probabilities*. Collectively, this initial information is known as *the priors*. The final answer - the

estimated probability that a patient has breast cancer, given that we know she has a positive result on her mammography - is known as the *revised probability* or the *posterior probability*. What we've just shown is that *if the two conditional probabilities are equal, the posterior probability equals the prior probability*.

---

**Fun Fact!****Q. How can I find the priors for a problem?**

- A. Many commonly used priors are listed in the *Handbook of Chemistry and Physics*.

**Q. Where do priors *originally* come from?**

- A. Never ask that question.

**Q. Uh huh. Then where do scientists get their priors?**

- A. Priors for scientific problems are established by annual vote of the AAAS. In recent years the vote has become fractious and controversial, with widespread acrimony, factional polarization, and several outright assassinations. This may be a front for infighting within the Bayes Council, or it may be that the disputants have too much spare time. No one is really sure.

**Q. I see. And where does everyone else get their priors?**

- A. They download their priors from Kazaa.

**Q. What if the priors I want aren't available on Kazaa?**

- A. There's a small, cluttered antique shop in a back alley of San Francisco's Chinatown. *Don't ask about the bronze rat.*

Actually, priors are true or false just like the final answer - they reflect reality and can be judged by comparing them against reality. For example, if you think that 920 out of 10,000 women in a sample have breast cancer, and the actual number is 100 out of 10,000, then your priors are wrong. For our particular problem, the priors might have been established by three studies - a study on the case histories of women with breast cancer to see how many of them tested positive on a mammography, a study on women without breast cancer to see how many of them test positive on a mammography, and an epidemiological study on the prevalence of breast cancer in some specific demographic.

---

Suppose that a barrel contains many small plastic eggs. Some eggs are painted red and some are painted blue. 40% of the eggs in the bin contain pearls, and 60% contain nothing. 30% of eggs containing pearls are painted blue, and 10% of eggs containing nothing are painted blue. What is the probability that a blue egg contains a pearl? For this example the arithmetic is simple enough that you may be able to do it in your head, and I would suggest trying to do so.

A more compact way of specifying the problem:

- $p(\text{pearl}) = 40\%$
- $p(\text{blue}|\text{pearl}) = 30\%$
- $p(\text{blue}|\sim\text{pearl}) = 10\%$
- $p(\text{pearl}|\text{blue}) = ?$

“ $\sim$ ” is shorthand for “not”, so  $\sim\text{pearl}$  reads “not pearl”.

$\text{blue}|\text{pearl}$  is shorthand for “blue given pearl” or “the probability that an egg is painted blue, given that the egg contains a pearl”. One thing that’s confusing about this notation is that the order of implication is read right-to-left, as in Hebrew or Arabic.  $\text{blue}|\text{pearl}$  means “blue $\leftarrow$ -pearl”, the degree to which pearl-ness implies blue-ness, not the degree to which blue-ness implies pearl-ness. This is confusing, but it’s unfortunately the standard notation in probability theory.

Readers familiar with quantum mechanics will have already encountered this peculiarity; in quantum mechanics, for example,  $\langle d | c \rangle \langle c | b \rangle \langle b | a \rangle$  reads as “the probability that a particle at A goes to B, then to C, ending up at D”. To follow the particle, you move your eyes from right to left. Reading from left to right, “|” means “given”; reading from right to left, “|” means “implies” or “leads to”. Thus, moving your eyes from left to right,  $\text{blue} | \text{pearl}$  reads “blue given pearl” or “the probability that an egg is painted blue, given that the egg contains a pearl”. Moving your eyes from right to left,  $\text{blue} | \text{pearl}$  reads “pearl implies blue” or “the probability that an egg containing a pearl is painted blue”.

The item on the right side is what you *already know* or the *premise*, and the item on the left side is the *implication* or *conclusion*. If we have  $p(\text{blue} | \text{pearl}) = 30\%$ , and we *already know* that some egg contains a pearl, then we can *conclude* there is a 30% chance that the egg is painted blue. Thus, the final fact we’re looking for - “the chance that a blue egg contains a pearl” or “the probability that an egg contains a pearl, if we know the egg is painted blue” - reads  $p(\text{pearl} | \text{blue})$ .

Let’s return to the problem. We have that 40% of the eggs contain pearls, and 60% of the eggs contain nothing. 30% of the eggs containing pearls are painted blue, so 12% of the eggs altogether contain pearls and are painted blue. 10% of the eggs containing nothing are painted blue, so altogether 6% of the eggs contain nothing and are painted blue. A total of 18% of the eggs are painted blue, and a total of 12% of the eggs are painted blue and contain pearls, so the chance a blue egg contains a pearl is 12/18 or  $2/3$  or around 67%.

The applet below, courtesy of Christian Rovner, shows a graphic representation of this problem:  
 (Are you having trouble seeing this applet? Do you see an image of the applet rather than the applet itself? Try downloading an updated Java<sup>2</sup>.)

---

<sup>2.</sup> <http://www.java.com/en/index.jsp>

Looking at this applet, it's easier to see why the final answer depends on all three probabilities; it's the *differential pressure* between the two conditional probabilities,  $p(\text{blue}|\text{pearl})$  and  $p(\text{blue}|\sim\text{pearl})$ , that *slides* the prior probability  $p(\text{pearl})$  to the posterior probability  $p(\text{pearl}|\text{blue})$ .

As before, we can see the necessity of all three pieces of information by considering extreme cases (feel free to type them into the applet). In a (large) barrel in which only one egg out of a thousand contains a pearl, knowing that an egg is painted blue slides the probability from 0.1% to 0.3% (instead of sliding the probability from 40% to 67%). Similarly, if 999 out of 1000 eggs contain pearls, knowing that an egg is blue slides the probability from 99.9% to 99.966%; the probability that the egg does *not* contain a pearl goes from 1/1000 to around 1/3000. Even when the prior probability changes, the differential pressure of the two conditional probabilities always slides the probability in the same *direction*. If you learn the egg is painted blue, the probability the egg contains a pearl always goes *up* - but it goes up *from* the prior probability, so you need to know the prior probability in order to calculate the final answer. 0.1% goes up to 0.3%, 10% goes up to 25%, 40% goes up to 67%, 80% goes up to 92%, and 99.9% goes up to 99.966%. If you're interested in knowing how any other probabilities slide, you can type your own prior probability into the Java applet. You can also click and drag the dividing line between `pearl` and `~pearl` in the upper bar, and watch the posterior probability change in the bottom bar.

Studies of clinical reasoning show that most doctors carry out the mental operation of *replacing* the original 1% probability with the 80% probability that a woman with cancer would get a positive mammography. Similarly, on the pearl-egg problem, most respondents unfamiliar with Bayesian reasoning would probably respond that the probability a blue egg contains a pearl is 30%, or perhaps 20% (the 30% chance of a true positive minus the 10% chance of a false positive). Even if this mental operation seems like a good idea at the time, it makes no sense in terms of the

question asked. It's like the experiment in which you ask a second-grader: "If eighteen people get on a bus, and then seven more people get on the bus, how old is the bus driver?" Many second-graders will respond: "Twenty-five." They understand when they're being prompted to carry out a particular mental procedure, but they haven't quite connected the procedure to reality. Similarly, to find the probability that a woman with a positive mammography has breast cancer, it makes no sense whatsoever to *replace* the original probability that the woman has cancer with the probability that a woman with breast cancer gets a positive mammography. Neither can you subtract the probability of a false positive from the probability of the true positive. These operations are as wildly irrelevant as adding the number of people on the bus to find the age of the bus driver.

---

I keep emphasizing the idea that evidence *slides* probability because of research that shows people tend to use spatial intuitions to grasp numbers. In particular, there's interesting evidence that we have an innate sense of quantity that's localized to left inferior parietal cortex - patients with damage to this area can selectively lose their sense of whether 5 is less than 8, while retaining their ability to read, write, and so on. (Yes, really!) The parietal cortex processes our sense of where things are in space (roughly speaking), so an innate "number line", or rather "quantity line", may be responsible for the human sense of numbers. This is why I suggest visualizing Bayesian evidence as *sliding* the probability along the number line; my hope is that this will translate Bayesian reasoning into something that makes sense to innate human brainware. (That, really, is what an "intuitive explanation" is.) For more information, see Stanislas Dehaene's *The Number Sense*.

---

A study by Gigerenzer and Hoffrage in 1995 showed that some ways of phrasing story problems are much more evocative of correct Bayesian reasoning. The *least* evocative phrasing used probabilities. A slightly more evocative phrasing used frequencies instead of probabilities; the problem remained the same, but instead of saying that 1% of women had breast cancer, one would

say that 1 out of 100 women had breast cancer, that 80 out of 100 women with breast cancer would get a positive mammography, and so on. Why did a higher proportion of subjects display Bayesian reasoning on this problem? Probably because saying “1 out of 100 women” encourages you to concretely visualize X women with cancer, leading you to visualize X women with cancer and a positive mammography, etc.

The most effective presentation found so far is what's known as *natural frequencies* - saying that 40 out of 100 eggs contain pearls, 12 out of 40 eggs containing pearls are painted blue, and 6 out of 60 eggs containing nothing are painted blue. A *natural frequencies* presentation is one in which the information about the prior probability is included in presenting the conditional probabilities. If you were just learning about the eggs' conditional probabilities through natural experimentation, you would - in the course of cracking open a hundred eggs - crack open around 40 eggs containing pearls, of which 12 eggs would be painted blue, while cracking open 60 eggs containing nothing, of which about 6 would be painted blue. In the course of learning the conditional probabilities, you'd see examples of blue eggs containing pearls about twice as often as you saw examples of blue eggs containing nothing.

It may seem like presenting the problem in this way is “cheating”, and indeed if it were a story problem in a math book, it probably *would* be cheating. However, if you're talking about real doctors, you *want* to cheat; you *want* the doctors to draw the right conclusions as easily as possible. The obvious next move would be to present all medical statistics in terms of natural frequencies. Unfortunately, while natural frequencies are a step in the right direction, it probably won't be enough. When problems are presented in natural frequencies, the proportion of people using Bayesian reasoning rises to around half. A big improvement, but not big enough when you're talking about real doctors and real patients.

A presentation of the problem in *natural frequencies* might be visualized like this:

In the frequency visualization, the *selective attrition* of the two conditional probabilities changes the *proportion* of eggs that contain pearls. The bottom bar is shorter than the top bar, just as the number of eggs painted blue is less than the total number of eggs. The probability graph shown earlier is really just the frequency graph with the bottom bar “renormalized”, stretched out to the same length as the top bar. In the frequency applet you can change the conditional probabilities by clicking and dragging the left and right edges of the graph. (For example, to change the conditional probability `blue | pearl`, click and drag the line on the left that stretches from the left edge of the top bar to the left edge of the bottom bar.)

In the probability applet, you can see that when the conditional probabilities are equal, there's no *differential* pressure - the arrows are the same size - so the prior probability doesn't slide between the top bar and the bottom bar. But the bottom bar in the probability applet is just a renormalized (stretched out) version of the bottom bar in the frequency applet, and the frequency applet shows *why* the probability doesn't slide if the two conditional probabilities are equal. Here's a case where the prior proportion of pearls remains 40%, and the proportion of pearl eggs painted blue remains 30%, but the number of empty eggs painted blue is also 30%:

If you diminish two shapes by the same factor, their relative proportion will be the same as before. If you diminish the left section of the top bar by the same factor as the right section, then the bottom bar will have the same proportions as the top bar - it'll just be smaller. If the two conditional probabilities are equal, learning that the egg is blue doesn't change the probability that the egg contains a pearl - for the same reason that similar triangles have identical angles; geometric figures don't change shape when you shrink them by a constant factor.

In this case, you might as well just say that *30% of eggs are painted blue*, since the probability of an egg being painted blue is independent of whether the egg contains a pearl. Applying a “test” that is statistically independent of its condition just shrinks the sample size. In this case, requiring that the egg be painted blue doesn’t shrink the group of eggs with pearls any more or less than it shrinks the group of eggs without pearls. It just shrinks the total number of eggs in the sample.

---

**Fun Fact!**

**Q. Why did the Bayesian reasoner cross the road?**

**A.** You need more information to answer this question.

---

Here’s what the original medical problem looks like when graphed. 1% of women have breast cancer, 80% of those women test positive on a mammography, and 9.6% of women without breast cancer also receive positive mammographies.

As is now clearly visible, the mammography doesn’t increase the probability a positive-testing woman has breast cancer by increasing the number of women with breast cancer - of course not; if mammography increased the number of women with breast cancer, no one would ever take the test! However, *requiring* a positive mammography is a membership test that *eliminates* many more women without breast cancer than women with cancer. The number of women without breast cancer diminishes by a factor of more than ten, from 9,900 to 950, while the number of women with breast cancer is diminished only from 100 to 80. Thus, the proportion of 80 within 1,030 is much larger than the proportion of 100 within 10,000. In the graph, the left sector (representing women with breast cancer) is small, but the mammography test projects almost all of this sector into the bottom bar. The right

sector (representing women without breast cancer) is large, but the mammography test projects a much smaller fraction of this sector into the bottom bar. There are, indeed, fewer women with breast cancer and positive mammographies than there are women with breast cancer - obeying the law of probabilities which requires that  $p(A) \geq p(A \& B)$ . But even though the left sector in the bottom bar is actually slightly smaller, the proportion of the left sector *within* the bottom bar is greater - though still not very great. If the bottom bar were renormalized to the same length as the top bar, it would look like the left sector had expanded. This is why the proportion of "women with breast cancer" in the group "women with positive mammographies" is higher than the proportion of "women with breast cancer" in the general population - although the proportion is still not very high. The evidence of the positive mammography slides the prior probability of 1% to the posterior probability of 7.8%.

---

Suppose there's yet another variant of the mammography test, mammography@, which behaves as follows. 1% of women in a certain demographic have breast cancer. Like ordinary mammography, mammography@ returns positive 9.6% of the time for women without breast cancer. However, mammography@ returns positive 0% of the time (say, once in a billion) for women with breast cancer. The graph for this scenario looks like this:

What is it that this test actually does? If a patient comes to you with a positive result on her mammography@, what do you say?

---

"Congratulations, you're among the rare 9.5% of the population whose health is definitely established by this test."

Mammography@ isn't a cancer test; it's a health test! Few women without breast cancer get positive results on mammography@, but *only* women without breast cancer ever get positive results at all. Not much of the right sector of the top bar projects into the

bottom bar, but *none* of the left sector projects into the bottom bar. So a positive result on mammography@ means you *definitely* don't have breast cancer.

---

What makes ordinary mammography a *positive* indicator for breast cancer is not that someone *named* the result "positive", but rather that the test result stands in a specific Bayesian relation to the condition of breast cancer. You could call the same result "positive" or "negative" or "blue" or "red" or "James Rutherford", or give it no name at all, and the test result would still slide the probability in exactly the same way. To minimize confusion, a test result which slides the probability of breast cancer upward should be called "positive". A test result which slides the probability of breast cancer downward should be called "negative". If the test result is statistically unrelated to the presence or absence of breast cancer - if the two conditional probabilities are equal - then we shouldn't call the procedure a "cancer test"! The *meaning* of the test is determined by the two conditional probabilities; any names attached to the results are simply convenient labels.

---

The bottom bar for the graph of mammography@ is small; mammography@ is a test that's only rarely useful. Or rather, the test only rarely gives *strong* evidence, and most of the time gives *weak* evidence. A negative result on mammography@ does slide probability - it just doesn't slide it very far. Click the "Result" switch at the bottom left corner of the applet to see what a *negative* result on mammography@ would imply. You might intuit that since the test *could* have returned positive for health, but didn't, then the failure of the test to return positive must mean that the woman has a higher chance of having breast cancer - that her probability of having breast cancer must be slid upward by the negative result on her health test.

This intuition is correct! The sum of the groups with negative results and positive results must always equal the group of all

women. If the positive-testing group has “more than its fair share” of women *without* breast cancer, there must be an at least slightly higher proportion of women *with* cancer in the negative-testing group. A positive result is rare but very strong evidence in one direction, while a negative result is common but very weak evidence in the opposite direction. You might call this the Law of Conservation of Probability - not a standard term, but the conservation rule is exact. If you take the revised probability of breast cancer after a positive result, times the *probability* of a positive result, and add that to the revised probability of breast cancer after a negative result, times the *probability* of a negative result, then you must always arrive at the prior probability. If you don’t yet *know* what the test result is, the *expected revised probability* after the test result arrives - taking both possible results into account - should always equal the prior probability.

On ordinary mammography, the test is expected to return “positive” 10.3% of the time - 80 positive women with cancer plus 950 positive women without cancer equals 1030 women with positive results. Conversely, the mammography should return negative 89.7% of the time:  $100\% - 10.3\% = 89.7\%$ . A positive result slides the revised probability from 1% to 7.8%, while a negative result slides the revised probability from 1% to 0.22%. So  $p(\text{cancer}|\text{positive}) * p(\text{positive}) + p(\text{cancer}|\text{negative}) * p(\text{negative}) = 7.8\% * 10.3\% + 0.22\% * 89.7\% = 1\% = p(\text{cancer})$ , as expected.

Why “as expected”? Let’s take a look at the quantities involved:

$p(\text{cancer}) :$	0.01	Group 1: 100 women with breast cancer
$p(\sim\text{cancer}) :$	0.99	Group 2: 9900 women without breast cancer
$p(\text{positive} \text{cancer}) :$	80.0%	80% of women with breast cancer have positive mammographies

$p(\text{~positive}   \text{cancer}) :$	20.0%	20% of women with breast cancer have negative mammographies
$p(\text{positive}   \text{~cancer}) :$	9.6%	9.6% of women without breast cancer have positive mammographies
$p(\text{~positive}   \text{~cancer}) :$	90.4%	90.4% of women without breast cancer have negative mammographies
$p(\text{cancer} \& \text{positive}) :$	0.008	Group A: 80 women with breast cancer and positive mammographies
$p(\text{cancer} \& \text{~positive}) :$	0.002	Group B: 20 women with breast cancer and negative mammographies
$p(\text{~cancer} \& \text{positive}) :$	0.095	Group C: 950 women without breast cancer and positive mammographies
$p(\text{~cancer} \& \text{~positive}) :$	0.895	Group D: 8950 women without breast cancer and negative mammographies
$p(\text{positive}) :$	0.103	1030 women with positive results
$p(\text{~positive}) :$	0.897	8970 women with negative results
$p(\text{cancer}   \text{positive}) :$	7.80%	Chance you have breast cancer if mammography is positive: 7.8%

$p(\sim \text{cancer}   \text{positive}) :$	92.20%	Chance you are healthy if mammography is positive: 92.2%
$p(\text{cancer}   \sim \text{positive}) :$	0.22%	Chance you have breast cancer if mammography is negative: 0.22%
$p(\sim \text{cancer}   \sim \text{positive}) :$	99.78%	Chance you are healthy if mammography is negative: 99.78%

One of the common confusions in using Bayesian reasoning is to mix up some or all of these quantities - which, as you can see, are all numerically different and have different meanings.  $p(A \& B)$  is the same as  $p(B \& A)$ , but  $p(A | B)$  is not the same thing as  $p(B | A)$ , and  $p(A \& B)$  is completely different from  $p(A | B)$ . (I don't know who chose the symmetrical " | " symbol to mean "implies", and then made the direction of implication right-to-left, but it was probably a bad idea.)

To get acquainted with all these quantities and the relationships between them, we'll play "follow the degrees of freedom". For example, the two quantities  $p(\text{cancer})$  and  $p(\sim \text{cancer})$  have 1 degree of freedom between them, because of the general law  $p(A) + p(\sim A) = 1$ . If you know that  $p(\sim \text{cancer}) = .99$ , you can obtain  $p(\text{cancer}) = 1 - p(\sim \text{cancer}) = .01$ . There's no room to say that  $p(\sim \text{cancer}) = .99$  and then also specify  $p(\text{cancer}) = .25$ ; it would violate the rule  $p(A) + p(\sim A) = 1$ .

$p(\text{positive} | \text{cancer})$  and  $p(\sim \text{positive} | \text{cancer})$  also have only one degree of freedom between them; either a woman with breast cancer gets a positive mammography or she doesn't. On the other hand,  $p(\text{positive} | \text{cancer})$  and  $p(\text{positive} | \sim \text{cancer})$  have *two* degrees of freedom. You can have a mammography test that returns positive for 80% of cancerous patients and 9.6% of healthy patients, or that returns positive for 70% of cancerous patients and 2% of healthy patients, or even a health test that returns "positive" for 30% of cancerous

patients and 92% of healthy patients. The two quantities, the output of the mammography test for cancerous patients and the output of the mammography test for healthy patients, are in mathematical terms independent; one cannot be obtained from the other in any way, and so they have two degrees of freedom between them.

What about  $p(\text{positive} \& \text{cancer})$ ,  $p(\text{positive} | \text{cancer})$ , and  $p(\text{cancer})$ ? Here we have three quantities; how many degrees of freedom are there? In this case the equation that must hold is  $p(\text{positive} \& \text{cancer}) = p(\text{positive} | \text{cancer}) * p(\text{cancer})$ . This equality reduces the degrees of freedom by one. If we know the fraction of patients with cancer, and chance that a cancerous patient has a positive mammography, we can deduce the fraction of patients who have breast cancer *and a* positive mammography by multiplying. You should recognize this operation from the graph; it's the projection of the top bar into the bottom bar.  $p(\text{cancer})$  is the left sector of the top bar, and  $p(\text{positive} | \text{cancer})$  determines how much of that sector projects into the bottom bar, and the left sector of the bottom bar is  $p(\text{positive} \& \text{cancer})$ .

Similarly, if we know the number of patients with breast cancer and positive mammographies, and also the number of patients with breast cancer, we can estimate the chance that a woman with breast cancer gets a positive mammography by dividing:  
 $p(\text{positive} | \text{cancer}) = p(\text{positive} \& \text{cancer}) / p(\text{cancer})$ . In fact, this is exactly how such medical diagnostic tests are calibrated; you do a study on 8,520 women with breast cancer and see that there are 6,816 (or thereabouts) women with breast cancer *and* positive mammographies, then divide 6,816 by 8520 to find that 80% of women with breast cancer had positive mammographies. (Incidentally, if you accidentally divide 8520 by 6,816 instead of the other way around, your calculations will start doing strange things, such as insisting that 125% of women with breast cancer and positive mammographies have breast cancer. This is a common mistake in carrying out Bayesian arithmetic, in

my experience.) And finally, if you know  $p(\text{positive} \& \text{cancer})$  and  $p(\text{positive} | \text{cancer})$ , you can deduce how many cancer patients there must have been originally. There are two degrees of freedom shared out among the three quantities; if we know any two, we can deduce the third.

How about  $p(\text{positive})$ ,  $p(\text{positive} \& \text{cancer})$ , and  $p(\text{positive} \& \sim \text{cancer})$ ? Again there are only two degrees of freedom among these three variables. The equation occupying the extra degree of freedom is  $p(\text{positive}) = p(\text{positive} \& \text{cancer}) + p(\text{positive} \& \sim \text{cancer})$ . This is how  $p(\text{positive})$  is computed to begin with; we figure out the number of women with breast cancer who have positive mammographies, and the number of women without breast cancer who have positive mammographies, then add them together to get the total number of women with positive mammographies. It would be very strange to go out and conduct a study to determine the number of women with positive mammographies - just that one number and nothing else - but in theory you could do so. And if you then conducted another study and found the number of those women who had positive mammographies *and* breast cancer, you would also know the number of women with positive mammographies and *no* breast cancer - either a woman with a positive mammography has breast cancer or she doesn't. In general,  $p(A \& B) + p(A \& \sim B) = p(A)$ . Symmetrically,  $p(A \& B) + p(\sim A \& B) = p(B)$ .

What about  $p(\text{positive} \& \text{cancer})$ ,  $p(\text{positive} \& \sim \text{cancer})$ ,  $p(\sim \text{positive} \& \text{cancer})$ , and  $p(\sim \text{positive} \& \sim \text{cancer})$ ? You might at first be tempted to think that there are only two degrees of freedom for these four quantities - that you can, for example, get  $p(\text{positive} \& \sim \text{cancer})$  by multiplying  $p(\text{positive}) * p(\sim \text{cancer})$ , and thus that all four quantities can be found given only the two quantities  $p(\text{positive})$  and  $p(\text{cancer})$ . This is not the case!  $p(\text{positive} \& \sim \text{cancer}) = p(\text{positive}) * p(\sim \text{cancer})$  only if the two probabilities are *statistically independent* - if the chance that a woman has breast cancer has no bearing on whether she has a positive mammography. As you'll

recall, this amounts to requiring that the two conditional probabilities be equal to each other - a requirement which would eliminate one degree of freedom. If you remember that these four quantities are the groups A, B, C, and D, you can look over those four groups and realize that, in theory, you can put any number of people into the four groups. If you start with a group of 80 women with breast cancer and positive mammographies, there's no reason why you can't add another group of 500 women with breast cancer and negative mammographies, followed by a group of 3 women without breast cancer and negative mammographies, and so on. So now it seems like the four quantities have four degrees of freedom. And they would, except that in expressing them as *probabilities*, we need to normalize them to *fractions* of the complete group, which adds the constraint that

$$p(\text{positive} \& \text{cancer}) + p(\text{positive} \& \sim \text{cancer}) + p(\sim \text{positive} \& \text{cancer}) + p(\sim \text{positive} \& \sim \text{cancer}) = 1.$$

This equation takes up one degree of freedom, leaving three degrees of freedom among the four quantities. If you specify the *fractions* of women in groups A, B, and D, you can deduce the fraction of women in group C.

---

Given the four groups A, B, C, and D, it is very straightforward to compute everything else:  $p(\text{cancer}) = A + B$ ,  $p(\sim \text{positive} \mid \text{cancer}) = B / (A + B)$ , and so on. Since ABCD contains three degrees of freedom, it follows that the entire set of 16 probabilities contains only three degrees of freedom. Remember that in our problems we always needed *three* pieces of information - the prior probability and the two conditional probabilities - which, indeed, have three degrees of freedom among them. Actually, for Bayesian problems, *any* three quantities with three degrees of freedom between them should logically specify the entire problem. For example, let's take a barrel of eggs with  $p(\text{blue}) = 0.40$ ,  $p(\text{blue} \mid \text{pearl}) = 5/13$ , and  $p(\sim \text{blue} \& \sim \text{pearl}) = 0.20$ . Given this information, you *can* compute  $p(\text{pearl} \mid \text{blue})$ .

As a story problem:

Suppose you have a large barrel containing a number of plastic

eggs. Some eggs contain pearls, the rest contain nothing. Some eggs are painted blue, the rest are painted red. Suppose that 40% of the eggs are painted blue,  $5/13$  of the eggs containing pearls are painted blue, and 20% of the eggs are both empty and painted red. What is the probability that an egg painted blue contains a pearl?

Try it - I assure you it is possible.

You probably shouldn't try to solve this with just a Javascript calculator, though. I used a Python console. (In theory, pencil and paper should also work, but I don't know anyone who owns a pencil so I couldn't try it personally.)

As a check on your calculations, does the (meaningless) quantity  $p(\sim\text{pearl} \mid \sim\text{blue}) / p(\text{pearl})$  roughly equal .51? (In story problem terms: The likelihood that a red egg is empty, divided by the likelihood that an egg contains a pearl, equals approximately .51.) Of course, using this information in the problem would be cheating.

If you can solve *that* problem, then when we revisit Conservation of Probability, it seems perfectly straightforward. Of course the mean revised probability, after administering the test, must be the same as the prior probability. Of course strong but rare evidence in one direction must be counterbalanced by common but weak evidence in the other direction.

Because:

$$\begin{aligned} & p(\text{cancer} \mid \text{positive}) * p(\text{positive}) \\ & + p(\text{cancer} \mid \sim\text{positive}) * p(\sim\text{positive}) \\ & = p(\text{cancer}) \end{aligned}$$

In terms of the four groups:

$$\begin{aligned} p(\text{cancer} \mid \text{positive}) &= A / (A + C) \\ p(\text{positive}) &= A + C \\ p(\text{cancer} \& \text{positive}) &= A \end{aligned}$$

$$\begin{aligned}
 p(\text{cancer} | \sim\text{positive}) &= B / (B + D) \\
 p(\sim\text{positive}) &= B + D \\
 p(\text{cancer} \& \sim\text{positive}) &= B \\
 p(\text{cancer}) &= A + B
 \end{aligned}$$


---

Let's return to the original barrel of eggs - 40% of the eggs containing pearls, 30% of the pearl eggs painted blue, 10% of the empty eggs painted blue. The graph for this problem is:

What happens to the revised probability,  $p(\text{pearl} | \text{blue})$ , if the proportion of eggs containing pearls is kept constant, but 60% of the eggs with pearls are painted blue (instead of 30%), and 20% of the empty eggs are painted blue (instead of 10%)? You could type 60% and 20% into the inputs for the two conditional probabilities, and see how the graph changes - but can you figure out in advance what the change will look like?

---

If you guessed that the revised probability *remains the same*, because the bottom bar grows by a factor of 2 but retains the same proportions, congratulations! Take a moment to think about how far you've come. Looking at a problem like

1% of women have breast cancer. 80% of women with breast cancer get positive mammographies. 9.6% of women without breast cancer get positive mammographies. If a woman has a positive mammography, what is the probability she has breast cancer?

the vast majority of respondents intuit that around 70-80% of women with positive mammographies have breast cancer. Now, looking at a problem like

Suppose there are two barrels containing many small plastic eggs. In both barrels, some eggs are painted blue and the

rest are painted red. In both barrels, 40% of the eggs contain pearls and the rest are empty. In the first barrel, 30% of the pearl eggs are painted blue, and 10% of the empty eggs are painted blue. In the second barrel, 60% of the pearl eggs are painted blue, and 20% of the empty eggs are painted blue. Would you rather have a blue egg from the first or second barrel?

you can see it's *intuitively obvious* that the probability of a blue egg containing a pearl is the same for either barrel. Imagine how hard it would be to see that using the old way of thinking!

---

It's intuitively obvious, but how to prove it? Suppose that we call P the prior probability that an egg contains a pearl, that we call M the first conditional probability (that a pearl egg is painted blue), and N the second conditional probability (that an empty egg is painted blue). Suppose that M and N are both increased or diminished by an arbitrary factor X - for example, in the problem above, they are both increased by a factor of 2. Does the revised probability that an egg contains a pearl, given that we know the egg is blue, stay the same?

- $p(\text{pearl}) = P$
- $p(\text{blue}|\text{pearl}) = M \cdot X$
- $p(\text{blue}|\sim\text{pearl}) = N \cdot X$
- $p(\text{pearl}|\text{blue}) = ?$

From these quantities, we get the four groups:

- Group A:  $p(\text{pearl}\&\text{blue}) = P \cdot M \cdot X$
- Group B:  $p(\text{pearl}\&\sim\text{blue}) = P \cdot (1 - (M \cdot X))$
- Group C:  $p(\sim\text{pearl}\&\text{blue}) = (1 - P) \cdot N \cdot X$
- Group D:  $p(\sim\text{pearl}\&\sim\text{blue}) = (1 - P) \cdot (1 - (N \cdot X))$

The proportion of eggs that contain pearls and are blue, within the group of all blue eggs, is then the proportion of group (A) within the group (A + C), equalling  $P \cdot M \cdot X / (P \cdot M \cdot X + (1 - P) \cdot N \cdot X)$ . The factor X in the numerator and denominator cancels out, so increasing or diminishing both conditional probabilities by a constant factor doesn't change the revised

probability.

---

**Fun  
Fact!**

**Q. Suppose that there are two barrels, each containing a number of plastic eggs. In both barrels, some eggs are painted blue and the rest are painted red. In the first barrel, 90% of the eggs contain pearls and 20% of the pearl eggs are painted blue. In the second barrel, 45% of the eggs contain pearls and 60% of the empty eggs are painted red. Would you rather have a blue pearl egg from the first or second barrel?**

**A.** Actually, it doesn't matter which barrel you choose! Can you see why?

---

*The probability that a test gives a true positive divided by the probability that a test gives a false positive* is known as the *likelihood ratio* of that test. Does the likelihood ratio of a medical test sum up everything there is to know about the usefulness of the test?

No, it does not! The likelihood ratio sums up everything there is to know about the *meaning* of a *positive* result on the medical test, but the meaning of a *negative* result on the test is not specified, nor is the frequency with which the test is useful. If we examine the algebra above, while  $p(\text{pearl}|\text{blue})$  remains constant,  $p(\text{pearl}|\sim\text{blue})$  may change - the X does *not* cancel out. As a story problem, this strange fact would look something like this:

Suppose that there are two barrels, each containing a number of plastic eggs. In both barrels, 40% of the eggs contain pearls and the rest contain nothing. In both barrels, some eggs are painted blue and the rest are painted red. In the first barrel, 30% of the eggs with pearls are painted blue, and 10% of the empty eggs are painted blue. In the second

barrel, 90% of the eggs with pearls are painted blue, and 30% of the empty eggs are painted blue. Would you rather have a blue egg from the first or second barrel? Would you rather have a red egg from the first or second barrel?

For the first question, the answer is that we don't care whether we get the blue egg from the first or second barrel. For the second question, however, the probabilities *do* change - in the first barrel, 34% of the red eggs contain pearls, while in the second barrel 8.7% of the red eggs contain pearls! Thus, we should prefer to get a red egg from the first barrel. In the first barrel, 70% of the pearl eggs are painted red, and 90% of the empty eggs are painted red. In the second barrel, 10% of the pearl eggs are painted red, and 70% of the empty eggs are painted red.

What goes on here? We start out by noting that, counter to intuition,  $p(\text{pearl}|\text{blue})$  and  $p(\text{pearl}|\sim\text{blue})$  have two degrees of freedom among them even when  $p(\text{pearl})$  is fixed - so there's no reason why one quantity shouldn't change while the other remains constant. But we didn't we just get through establishing a law for "Conservation of Probability", which says that  $p(\text{pearl}|\text{blue}) * p(\text{blue}) + p(\text{pearl}|\sim\text{blue}) * p(\sim\text{blue}) = p(\text{pearl})$ ? Doesn't this equation take up one degree of freedom? No, because  $p(\text{blue})$  isn't fixed between the two problems. In the second barrel, the proportion of blue eggs containing pearls is the same as in the first barrel, but a much larger fraction of eggs are painted blue! This alters the set of *red* eggs in such a way that the proportions *do* change. Here's a graph for the red eggs in the second barrel:

---

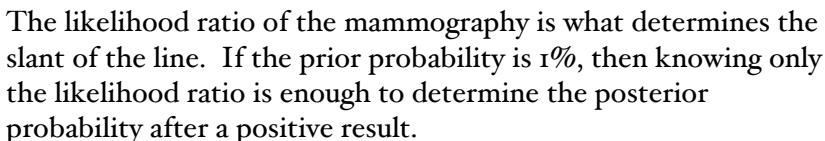


---

Let's return to the example of a medical test. The likelihood ratio of a medical test - the number of true positives divided by the number of false positives - tells us everything there is to know about the *meaning* of a *positive* result. But it doesn't tell us the meaning of a negative result, and it doesn't tell us how often the test is useful. For example, a mammography with a hit rate of 80%

for patients with breast cancer and a false positive rate of 9.6% for healthy patients has the same likelihood ratio as a test with an 8% hit rate and a false positive rate of 0.96%. Although these two tests have the same likelihood ratio, the first test is more useful in every way - it detects disease more often, and a negative result is stronger evidence of health.

The likelihood ratio for a positive result summarizes the differential pressure of the two conditional probabilities for a positive result, and thus summarizes how much a positive result will slide the prior probability. Take a probability graph, like this one:



The likelihood ratio of the mammography is what determines the slant of the line. If the prior probability is 1%, then knowing only the likelihood ratio is enough to determine the posterior probability after a positive result.

But, as you can see from the frequency graph, the likelihood ratio doesn't tell the whole story - in the frequency graph, the *proportions* of the bottom bar can stay fixed while the *size* of the bottom bar changes.  $p(\text{blue})$  increases but  $p(\text{pearl}|\text{blue})$  doesn't change, because  $p(\text{pearl} \& \text{blue})$  and  $p(\sim\text{pearl} \& \text{blue})$  increase by the same factor. But when you flip the graph to look at  $p(\sim\text{blue})$ , the proportions of  $p(\text{pearl} \& \sim\text{blue})$  and  $p(\sim\text{pearl} \& \sim\text{blue})$  do *not* remain constant.

Of course the likelihood ratio *can't* tell the whole story; the likelihood ratio and the prior probability together are only two numbers, while the problem has three degrees of freedom.

---

Suppose that you apply *two* tests for breast cancer in succession - say, a standard mammography and also some other test which is independent of mammography. Since I don't know of any such test which is *independent* of mammography, I'll invent one for the

purpose of this problem, and call it the Tams-Braylor Division Test, which checks to see if any cells are dividing more rapidly than other cells. We'll suppose that the Tams-Braylor gives a true positive for 90% of patients with breast cancer, and gives a false positive for 5% of patients without cancer. Let's say the prior prevalence of breast cancer is 1%. If a patient gets a positive result on her mammography *and* her Tams-Braylor, what is the revised probability she has breast cancer?

One way to solve this problem would be to take the revised probability for a positive mammography, which we already calculated as 7.8%, and plug that into the Tams-Braylor test as the new prior probability. If we do this, we find that the result comes out to 60%.

But this assumes that first we see the positive mammography result, and then the positive result on the Tams-Braylor. What if first the woman gets a positive result on the Tams-Braylor, followed by a positive result on her mammography. Intuitively, it seems like it shouldn't matter. Does the math check out?

First we'll administer the Tams-Braylor to a woman with a 1% prior probability of breast cancer.

Then we administer a mammography, which gives 80% true positives and 9.6% false positives, and it also comes out positive.

Lo and behold, the answer is again 60%. (If it's not exactly the same, it's due to rounding error - you can get a more precise calculator, or work out the fractions by hand, and the numbers will be exactly equal.)

An algebraic proof that both strategies are equivalent is left to the reader. To visualize, imagine that the lower bar of the frequency applet for mammography projects an even lower bar using the probabilities of the Tams-Braylor Test, and that the final lowest bar is the same regardless of the order in which the conditional probabilities are projected.

---

We might also reason that since the two tests are independent, the probability a woman with breast cancer gets a positive mammography *and* a positive Tams-Braylor is  $90\% * 80\% = 72\%$ . And the probability that a woman without breast cancer gets false positives on mammography and Tams-Braylor is  $5\% * 9.6\% = 0.48\%$ . So if we wrap it all up as a single test with a likelihood ratio of  $72\%/0.48\%$ , and apply it to a woman with a 1% prior probability of breast cancer:

...we find once again that the answer is 60%.

Suppose that the prior prevalence of breast cancer in a demographic is 1%. Suppose that we, as doctors, have a repertoire of three independent tests for breast cancer. Our first test, test A, a mammography, has a likelihood ratio of  $80\%/9.6\% = 8.33$ . The second test, test B, has a likelihood ratio of 18.0 (for example, from 90% versus 5%); and the third test, test C, has a likelihood ratio of 3.5 (which could be from 70% versus 20%, or from 35% versus 10%; it makes no difference). Suppose a patient gets a positive result on all three tests. What is the probability the patient has breast cancer?

Here's a fun trick for simplifying the bookkeeping. If the prior prevalence of breast cancer in a demographic is 1%, then 1 out of 100 women have breast cancer, and 99 out of 100 women do not have breast cancer. So if we rewrite the *probability* of 1% as an *odds ratio*, the odds are:

1 : 99

And the likelihood ratios of the three tests A, B, and C are:

$$\begin{aligned} 8.33:1 &= 25:3 \\ 18.0:1 &= 18:1 \\ 3.5:1 &= 7:2 \end{aligned}$$

The *odds* for women with breast cancer who score positive on all three tests, versus women without breast cancer who score

positive on all three tests, will equal:

$$1 \cdot 25 \cdot 18 \cdot 7 : 99 \cdot 3 \cdot 1 \cdot 2 = \\ 3,150 : 594$$

To recover the probability from the odds, we just write:  
 $3,150 / (3,150 + 594) = 84\%$

This always works regardless of how the odds ratios are written; i.e., 8.33:1 is just the same as 25:3 or 75:9. It doesn't matter in what order the tests are administered, or in what order the results are computed. The proof is left as an exercise for the reader.

---



---

E. T. Jaynes, in "Probability Theory With Applications in Science and Engineering", suggests that credibility and evidence should be measured in decibels.

Decibels?

Decibels are used for measuring exponential differences of intensity. For example, if the sound from an automobile horn carries 10,000 times as much energy (per square meter per second) as the sound from an alarm clock, the automobile horn would be 40 decibels louder. The sound of a bird singing might carry 1,000 times less energy than an alarm clock, and hence would be 30 decibels softer. To get the number of decibels, you take the logarithm base 10 and multiply by 10.

$$\text{decibels} = 10 \log_{10} (\text{intensity})$$

*or*

$$\text{intensity} = 10^{(\text{decibels}/10)}$$

Suppose we start with a prior probability of 1% that a woman has breast cancer, corresponding to an odds ratio of 1:99. And then we administer three tests of likelihood ratios 25:3, 18:1, and 7:2. You could multiply those numbers... or you could just add their logarithms:

10  $\log_{10} (1/99) = -20$   
10  $\log_{10} (25/3) = 9$   
10  $\log_{10} (18/1) = 13$   
10  $\log_{10} (7/2) = 5$

It starts out as fairly unlikely that a woman has breast cancer - our credibility level is at -20 decibels. Then three test results come in, corresponding to 9, 13, and 5 decibels of evidence. This raises the credibility level by a total of 27 decibels, meaning that the prior credibility of -20 decibels goes to a posterior credibility of 7 decibels. So the odds go from 1:99 to 5:1, and the probability goes from 1% to around 83%.

---

In front of you is a bookbag containing 1,000 poker chips. I started out with two such bookbags, one containing 700 red and 300 blue chips, the other containing 300 red and 700 blue. I flipped a fair coin to determine which bookbag to use, so your prior probability that the bookbag in front of you is the red bookbag is 50%. Now, you sample randomly, with replacement after each chip. In 12 samples, you get 8 reds and 4 blues. What is the probability that this is the predominantly red bag?

Just for fun, try and work this one out in your head. You don't need to be exact - a rough estimate is good enough. When you're ready, continue onward.

---

According to a study performed by Lawrence Phillips and Ward Edwards in 1966, most people, faced with this problem, give an answer in the range 70% to 80%. Did you give a substantially higher probability than that? If you did, congratulations - Ward Edwards wrote that very seldom does a person answer this question properly, even if the person is relatively familiar with Bayesian reasoning. The correct answer is 97%.

The likelihood ratio for the test result “red chip” is  $7/3$ , while the likelihood ratio for the test result “blue chip” is  $3/7$ . Therefore a blue chip is exactly the same amount of evidence as a red chip, just in the other direction - a red chip is 3.6 decibels of evidence for the red bag, and a blue chip is -3.6 decibels of evidence. If you draw one blue chip and one red chip, they cancel out. So the *ratio* of red chips to blue chips does not matter; only the *excess* of red chips over blue chips matters. There were eight red chips and four blue chips in twelve samples; therefore, four *more* red chips than blue chips. Thus the posterior odds will be:

$$7^4 : 3^4 = 2401 : 81$$

which is around 30:1, i.e., around 97%.

The prior credibility starts at 0 decibels and there's a total of around 14 decibels of evidence, and indeed this corresponds to odds of around 25:1 or around 96%. Again, there's some rounding error, but if you performed the operations using exact arithmetic, the results would be identical.

We can now see *intuitively* that the bookbag problem would have exactly the same answer, obtained in just the same way, if sixteen chips were sampled and we found ten red chips and six blue chips.

---

You are a mechanic for gizmos. When a gizmo stops working, it is due to a blocked hose 30% of the time. If a gizmo's hose is blocked, there is a 45% probability that prodding the gizmo will produce sparks. If a gizmo's hose is unblocked, there is only a 5% chance that prodding the gizmo will produce sparks. A customer brings you a malfunctioning gizmo. You prod the gizmo and find that it produces sparks. What is the probability that a spark-producing gizmo has a blocked hose?

What is the sequence of arithmetical operations that you performed to solve this problem?

$$(45\% * 30\%) / (45\% * 30\% + 5\% * 70\%)$$

Similarly, to find the chance that a woman with positive mammography has breast cancer, we computed:

$$\frac{p(\text{positive}|\text{cancer}) * p(\text{cancer})}{p(\text{positive}|\text{cancer}) * p(\text{cancer}) + p(\text{positive}|\sim\text{cancer}) * p(\sim\text{cancer})}$$

*which is*

$$\frac{p(\text{positive}\&\text{cancer})}{p(\text{positive}\&\text{cancer}) + p(\text{positive}\&\sim\text{cancer})}$$

*which is*

$$\frac{p(\text{positive}\&\text{cancer})}{p(\text{positive})}$$

*which is*

$$p(\text{cancer}|\text{positive})$$

The fully general form of this calculation is known as *Bayes' Theorem* or *Bayes' Rule*:

Bayes' Theorem:

$$p(A|X) = \frac{p(X|A) * p(A)}{p(X|A) * p(A) + p(X|\sim A) * p(\sim A)}$$

Given some phenomenon A that we want to investigate, and an observation X that is evidence about A - for example, in the previous example, A is breast cancer and X is a positive mammography - Bayes' Theorem tells us how we should *update* our probability of A, given the *new evidence* X.

By this point, Bayes' Theorem may seem blatantly obvious or even tautological, rather than exciting and new. If so, this introduction has *entirely succeeded* in its purpose.

---

**Fun Fact!****Q. Who originally discovered Bayes' Theorem?**

**A.** The Reverend Thomas Bayes, by far the most enigmatic figure in mathematical history. Almost nothing is known of Bayes's life, and very few of his manuscripts survived. Thomas Bayes was born in 1701 or 1702 to Joshua Bayes and Ann Carpenter, and his date of death is listed as 1761. The exact date of Thomas Bayes's birth is not known for certain because Joshua Bayes, though a surprisingly wealthy man, was a member of an unusual, esoteric, and even heretical religious sect, the "Nonconformists". The Nonconformists kept their birth registers secret, supposedly from fear of religious discrimination; whatever the reason, no true record exists of Thomas Bayes's birth. Thomas Bayes was raised a Nonconformist and was soon promoted into the higher ranks of the Nonconformist theosophers, whence comes the "Reverend" in his name.

In 1742 Bayes was elected a Fellow of the Royal Society of London, the most prestigious scientific body of its day, despite Bayes having published no scientific or mathematical works at that time. Bayes's nomination certificate was signed by sponsors including the President and the Secretary of the Society, making his

election almost certain. Even today, however, it remains a mystery *why* such weighty names sponsored an unknown into the Royal Society.

Bayes's sole publication during his known lifetime was allegedly a mystical book entitled *Divine Benevolence*, laying forth the original causation and ultimate purpose of the universe. The book is commonly attributed to Bayes, though it is said that no author appeared on the title page, and the entire work is sometimes considered to be of dubious provenance.

Most mysterious of all, Bayes' Theorem itself appears in a Bayes manuscript presented to the Royal Society of London in 1764, *three years after Bayes's supposed death in 1761!*

Despite the shocking circumstances of its presentation, Bayes' Theorem was soon forgotten, and was popularized within the scientific community only by the later efforts of the great mathematician Pierre-Simon Laplace. Laplace himself is almost as enigmatic as Bayes; we don't even know whether it was "Pierre" or "Simon" that was his actual first name. Laplace's papers are said to have contained a design for an AI capable of predicting all future events, the so-called "Laplacian superintelligence". While it is generally believed that Laplace never tried to implement his design, there remains the fact that Laplace presciently fled the guillotine that claimed many of his colleagues during the Reign of Terror.

Even today, physicists sometimes attribute unusual effects to a “Laplacian Operator” intervening in their experiments.

In summary, we do not know the real circumstances of Bayes's birth, the ultimate origins of Bayes' Theorem, Bayes's actual year of death, or even whether Bayes ever really died. Nonetheless “Reverend Thomas Bayes”, whatever his true identity, has the greatest fondness and gratitude of Earth's scientific community.

---

So why is it that some people are so *excited* about Bayes' Theorem?

“Do you believe that a nuclear war will occur in the next 20 years? If no, why not?” Since I wanted to use some common answers to this question to make a point about rationality, I went ahead and asked the above question in an IRC channel, #philosophy on EFNet.

One EFNetter who answered replied “No” to the above question, but added that he believed biological warfare would wipe out “99.4%” of humanity within the next ten years. I then asked whether he believed 100% was a possibility. “No,” he said. “Why not?”, I asked. “Because I'm an optimist,” he said. (Roanoke of #philosophy on EFNet wishes to be credited with this statement, even having been warned that it will not be cast in a complimentary light. Good for him!) Another person who answered the above question said that 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.

What is it *exactly* that makes these thoughts “irrational” - a poor way of arriving at truth? There are a number of intuitive replies

that can be given to this; for example: “It is not rational to believe things only because they are comforting.” Of course it is equally irrational to believe things only because they are *discomforting*; the second error is less common, but equally irrational. Other intuitive arguments include the idea that “Whether or not you happen to be an optimist has nothing to do with whether biological warfare wipes out the human species”, or “Pure hope is not evidence about nuclear war because it is not an observation about nuclear war.”

There is also a mathematical reply that is precise, exact, and contains all the intuitions as special cases. This mathematical reply is known as Bayes’ Theorem.

For example, the reply “Whether or not you happen to be an optimist has nothing to do with whether biological warfare wipes out the human species” can be translated into the statement:

$$\begin{aligned} p(\text{you are currently an optimist} \mid \text{biological war occurs within ten years and wipes out humanity}) &= \\ p(\text{you are currently an optimist} \mid \text{biological war occurs within ten years and does not wipe out humanity}) \end{aligned}$$

Since the two probabilities for  $p(X \mid A)$  and  $p(X \mid \sim A)$  are equal, Bayes’ Theorem says that  $p(A \mid X) = p(A)$ ; as we have earlier seen, when the two conditional probabilities are equal, the revised probability equals the prior probability. If X and A are unconnected - statistically independent - then finding that X is true cannot be evidence that A is true; observing X does not update our probability for A; saying “X” is not an argument for A.

But suppose you are arguing with someone who is verbally clever and who says something like, “Ah, but since I’m an optimist, I’ll have renewed hope for tomorrow, work a little harder at my dead-end job, pump up the global economy a little, eventually, through the trickle-down effect, sending a few dollars into the pocket of the researcher who ultimately finds a way to stop biological warfare - so you see, the two events are related after all, and I can use one as valid evidence about the other.” In one sense, this is

correct - *any* correlation, no matter how weak, is fair prey for Bayes' Theorem; *but* Bayes' Theorem distinguishes between weak and strong evidence. That is, Bayes' Theorem not only tells us what is and isn't evidence, it also describes the *strength* of evidence. Bayes' Theorem not only tells us *when* to revise our probabilities, but *how much* to revise our probabilities. A correlation between hope and biological warfare may exist, but it's a lot weaker than the speaker wants it to be; he is revising his probabilities much too far.

Let's say you're a woman who's just undergone a mammography. Previously, you figured that you had a very small chance of having breast cancer; we'll suppose that you read the statistics somewhere and so you know the chance is 1%. When the positive mammography comes in, your estimated chance should now shift to 7.8%. There is no room to say something like, "Oh, well, a positive mammography isn't definite evidence, some healthy women get positive mammographies too. I don't want to despair too early, and I'm not going to revise my probability until more evidence comes in. Why? Because I'm a optimist." And there is similarly no room for saying, "Well, a positive mammography may not be definite evidence, but I'm going to assume the worst until I find otherwise. Why? Because I'm a pessimist." Your revised probability should go to 7.8%, no more, no less.

Bayes' Theorem describes what makes something "evidence" and how much evidence it is. Statistical models are judged by comparison to the *Bayesian method* because, in statistics, the Bayesian method is as good as it gets - the Bayesian method defines the maximum amount of mileage you can get out of a given piece of evidence, in the same way that thermodynamics defines the maximum amount of work you can get out of a temperature differential. This is why you hear cognitive scientists talking about *Bayesian reasoners*. In cognitive science, *Bayesian reasoner* is the technically precise codeword that we use to mean *rational mind*.

There are also a number of general heuristics about human reasoning that you can learn from looking at Bayes' Theorem.

For example, in many discussions of Bayes' Theorem, you may hear cognitive psychologists saying that people *do not take prior frequencies sufficiently into account*, meaning that when people approach a problem where there's some evidence X indicating that condition A might hold true, they tend to judge A's likelihood solely by how well the evidence X seems to match A, without taking into account the prior frequency of A. If you think, for example, that under the mammography example, the woman's chance of having breast cancer is in the range of 70%-80%, then this kind of reasoning is insensitive to the prior frequency given in the problem; it doesn't notice whether 1% of women or 10% of women start out having breast cancer. "Pay more attention to the prior frequency!" is one of the many things that humans need to bear in mind to partially compensate for our built-in inadequacies.

A related error is to pay too much attention to  $p(X|A)$  and not enough to  $p(X|\neg A)$  when determining how much evidence X is for A. The degree to which a result X is *evidence for A* depends, not only on the strength of the statement *we'd expect to see result X if A were true*, but also on the strength of the statement *we wouldn't expect to see result X if A weren't true*. For example, if it is raining, this very strongly implies the grass is wet –  $p(\text{wetgrass}|\text{rain}) \sim 1$  – but seeing that the grass is wet doesn't necessarily mean that it has just rained; perhaps the sprinkler was turned on, or you're looking at the early morning dew. Since  $p(\text{wetgrass}|\neg\text{rain})$  is substantially greater than zero,  $p(\text{rain}|\text{wetgrass})$  is substantially less than one. On the other hand, if the grass was *never* wet when it wasn't raining, then knowing that the grass was wet would *always* show that it was raining,  $p(\text{rain}|\text{wetgrass}) \sim 1$ , even if  $p(\text{wetgrass}|\text{rain}) = 50\%$ ; that is, even if the grass only got wet 50% of the times it rained. Evidence is always the result of the *differential* between the two conditional probabilities. *Strong* evidence is not the product of a very high probability that A leads to X, but the product of a very *low* probability that *not-A* could have led to X.

The *Bayesian revolution in the sciences* is fueled, not only by more and more cognitive scientists suddenly noticing that mental

phenomena have Bayesian structure in them; not only by scientists in every field learning to judge their statistical methods by comparison with the Bayesian method; but also by the idea that *science itself is a special case of Bayes' Theorem; experimental evidence is Bayesian evidence.* The Bayesian revolutionaries hold that when you perform an experiment and get evidence that “confirms” or “disconfirms” your theory, this confirmation and disconfirmation is governed by the Bayesian rules. For example, you have to take into account, not only whether your theory predicts the phenomenon, but whether other possible explanations also predict the phenomenon. Previously, the most popular philosophy of science was probably Karl Popper's *falsificationism* - this is the old philosophy that the Bayesian revolution is currently dethroning. Karl Popper's idea that theories can be definitely falsified, but never definitely confirmed, is yet another special case of the Bayesian rules; if  $p(X|A) \sim 1$  - if the theory makes a definite prediction - then observing  $\neg X$  very strongly falsifies A. On the other hand, if  $p(X|A) \sim 1$ , and we observe X, this doesn't definitely confirm the theory; there might be some other condition B such that  $p(X|B) \sim 1$ , in which case observing X doesn't favor A over B. For observing X to definitely confirm A, we would have to know, not that  $p(X|A) \sim 1$ , but that  $p(X|\neg A) \sim 0$ , which is something that we can't know because we can't range over all possible alternative explanations. For example, when Einstein's theory of General Relativity toppled Newton's incredibly well-confirmed theory of gravity, it turned out that all of Newton's predictions were just a special case of Einstein's predictions.

You can even formalize Popper's philosophy mathematically. The likelihood ratio for X,  $p(X|A) / p(X|\neg A)$ , determines how much observing X slides the probability for A; the likelihood ratio is what says *how strong* X is as evidence. Well, in your theory A, you can predict X with probability 1, if you like; but you can't control the denominator of the likelihood ratio,  $p(X|\neg A)$  - there will always be some alternative theories that also predict X, and while we go with the simplest theory that fits the current evidence, you may someday encounter some evidence that an alternative theory predicts but your theory does not. That's the hidden gotcha that

toppled Newton's theory of gravity. So there's a limit on how much mileage you can get from successful predictions; there's a limit on how high the likelihood ratio goes for *confirmatory* evidence.

On the other hand, if you encounter some piece of evidence Y that is definitely *not* predicted by your theory, this is *enormously* strong evidence *against* your theory. If  $p(Y|A)$  is infinitesimal, then the likelihood ratio will also be infinitesimal. For example, if  $p(Y|A)$  is 0.0001%, and  $p(Y|\sim A)$  is 1%, then the likelihood ratio  $p(Y|A)/p(Y|\sim A)$  will be 1:10000. -40 decibels of evidence! Or flipping the likelihood ratio, if  $p(Y|A)$  is *very small*, then  $p(Y|\sim A)/p(Y|A)$  will be *very large*, meaning that observing Y greatly favors  $\sim A$  over A. Falsification is much stronger than confirmation. This is a consequence of the earlier point that *very strong* evidence is not the product of a very high probability that A leads to X, but the product of a very *low* probability that *not-A* could have led to X. This is the precise Bayesian rule that underlies the heuristic value of Popper's falsificationism.

Similarly, Popper's dictum that an idea must be falsifiable can be interpreted as a manifestation of the Bayesian conservation-of-probability rule; if a result X is positive evidence for the theory, then the result  $\sim X$  would have disconfirmed the theory to some extent. If you try to interpret both X and  $\sim X$  as "confirming" the theory, the Bayesian rules say this is impossible! To increase the probability of a theory you *must* expose it to tests that can potentially decrease its probability; this is not just a rule for detecting would-be cheaters in the social process of science, but a consequence of Bayesian probability theory. On the other hand, Popper's idea that there is *only* falsification and *no such thing* as confirmation turns out to be incorrect. Bayes' Theorem shows that falsification is *very strong* evidence compared to confirmation, but falsification is still probabilistic in nature; it is not governed by fundamentally different rules from confirmation, as Popper argued.

So we find that many phenomena in the cognitive sciences, plus the statistical methods used by scientists, plus the scientific

method itself, are all turning out to be special cases of Bayes' Theorem. Hence the Bayesian revolution.

---

**Fun Fact!**
**Q. Are there any limits to the power of Bayes' Theorem?**

**A.** According to legend, one who fully grasped Bayes' Theorem would gain the ability to create and physically enter an alternate universe using only off-the-shelf equipment and a short computer program. One who fully grasps Bayes' Theorem, yet remains in our universe to aid others, is known as a Bayesattva.

Bayes' Theorem:

$$p(A|X) = \frac{p(X|A) * p(A)}{p(X|A) * p(A) + p(X|\sim A) * p(\sim A)}$$

Why wait so long to introduce Bayes' Theorem, instead of just showing it at the beginning? Well... because I've tried that before; and what happens, in my experience, is that people get all tangled up in trying to apply Bayes' Theorem as a set of *poorly grounded mental rules*; instead of the Theorem helping, it becomes *one more thing to juggle mentally*, so that in addition to trying to remember how many women with breast cancer have positive mammographies, the reader is also trying to remember whether it's  $p(X|A)$  in the numerator or  $p(A|X)$ , and whether a positive mammography result corresponds to A or X, and which side of  $p(X|A)$  is the implication, and what the terms are in the denominator, and so on. In this excruciatingly gentle introduction, I tried to show all the workings of Bayesian reasoning *without* ever introducing the explicit Theorem as

something extra to memorize, hopefully reducing the number of factors the reader needed to mentally juggle.

Even if you happen to be one of the fortunate people who can easily grasp and apply abstract theorems, the mental-juggling problem is still something to bear in mind if you ever need to explain Bayesian reasoning to someone else.

If you do find yourself losing track, my advice is to forget Bayes' Theorem as an *equation* and think about the *graph*.  $p(A)$  and  $p(\neg A)$  are at the top.  $p(X|A)$  and  $p(X|\neg A)$  are the projection factors.  $p(X \& A)$  and  $p(X \& \neg A)$  are at the bottom. And  $p(A|X)$  equals the proportion of  $p(X \& A)$  within  $p(X \& A) + p(X \& \neg A)$ . The graph isn't shown here - but can you see it in your mind?

And if thinking about the graph doesn't work, I suggest forgetting about Bayes' Theorem entirely - just try to work out the specific problem in gizmos, hoses, and sparks, or whatever it is.

---

Having introduced Bayes' Theorem explicitly, we can explicitly discuss its components.

$$p(A|X) = \frac{p(X|A) * p(A)}{p(X|A) * p(A) + p(X|\neg A) * p(\neg A)}$$

We'll start with  $p(A|X)$ . If you ever find yourself getting confused about what's A and what's X in Bayes' Theorem, start with  $p(A|X)$  on the left side of the equation; that's the simplest part to interpret. A is the thing we want to know about. X is how we're observing it; X is the evidence we're using to make inferences about A. Remember that for every expression  $p(Q|P)$ , we want to know about the probability for Q given P, the degree to which P implies Q - a more sensible notation, which it is now too late to adopt, would be  $p(Q <- P)$ .

$p(Q|P)$  is closely related to  $p(Q \& P)$ , but they are not identical. Expressed as a probability or a fraction,  $p(Q \& P)$  is the proportion of things that have property Q and property P within *all things*; i.e.,

the proportion of “women with breast cancer and a positive mammography” within the group of *all women*. If the total number of women is 10,000, and 80 women have breast cancer and a positive mammography, then  $p(Q \& P)$  is  $80/10,000 = 0.8\%$ . You might say that the absolute quantity, 80, is being normalized to a probability relative to the *group of all women*. Or to make it clearer, suppose that there’s a group of 641 women with breast cancer and a positive mammography within a total sample group of 89,031 women. 641 is the absolute quantity. If you pick out a random woman from the *entire sample*, then the *probability* you’ll pick a woman with breast cancer and a positive mammography is  $p(Q \& P)$ , or 0.72% (in this example).

On the other hand,  $p(Q|P)$  is the proportion of things that have property Q and property P within *all things that have P*; i.e., the proportion of women with breast cancer and a positive mammography within the group of *all women with positive mammographies*. If there are 641 women with breast cancer and positive mammographies, 7915 women with positive mammographies, and 89,031 women, then  $p(Q \& P)$  is the probability of getting one of those 641 women if you’re picking at random from the entire group of 89,031, while  $p(Q|P)$  is the probability of getting one of those 641 women if you’re picking at random from the smaller group of 7915.

In a sense,  $p(Q | P)$  really means  $p(Q \& P | P)$ , but specifying the extra P all the time would be redundant. You already *know* it has property P, so the property you’re *investigating* is Q - even though you’re looking at the size of group Q&P within group P, not the size of group Q within group P (which would be nonsense). This is what it means to take the property on the right-hand side as *given*; it means you know you’re working only within the group of things that have property P. When you constrict your focus of attention to see only this smaller group, many other probabilities change. If you’re taking P as *given*, then  $p(Q \& P)$  equals just  $p(Q)$  - at least, *relative to the group P*. The *old*  $p(Q)$ , the frequency of “things that have property Q within the entire sample”, is revised to the new frequency of “things that have property Q within the subsample of things that have property P”. If P is *given*, if P is our entire world,

then looking for Q&P is the same as looking for just Q.

If you constrict your focus of attention to only the population of eggs that are painted blue, then suddenly “the probability that an egg contains a pearl” becomes a different number; this proportion is different for the population of blue eggs than the population of all eggs. The *given*, the property that constricts our focus of attention, is always on the *right* side of  $p(Q|P)$ ; the P becomes our world, the entire thing we see, and on the other side of the “given” P always has probability 1 - that is what it means to take P as given. So  $p(Q|P)$  means “If P has probability 1, what is the probability of Q?” or “If we constrict our attention to only things or events where P is true, what is the probability of Q?” Q, on the other side of the given, is *not* certain - its probability may be 10% or 90% or any other number. So when you use Bayes’ Theorem, and you write the part on the left side as  $p(A|X)$  - how to *update* the probability of A after seeing X, the new probability of A *given* that we know X, the degree to which X *implies* A - you can tell that X is always the *observation* or the *evidence*, and A is the property being investigated, the thing you want to know about.

---

The right side of Bayes’ Theorem is derived from the left side through these steps:

$$\begin{aligned} p(A|X) &= \frac{p(A|X)}{p(X\&A)} \\ p(A|X) &= \frac{p(X\&A)}{p(X)} \\ p(A|X) &= \frac{p(X\&A)}{p(X\&A) + p(X\&\sim A)} \\ p(A|X) &= \frac{p(X|A) * p(A)}{p(X|A) * p(A) + p(X|\sim A) * p(\sim A)} \end{aligned}$$

The first step,  $p(A|X)$  to  $p(X\&A) / p(X)$ , may look like a tautology. The actual math performed is different, though.  $p(A|X)$  is a single number, the normalized probability or frequency of A within the subgroup X.  $p(X\&A) / p(X)$  are usually the percentage frequencies of X&A and X within the entire

sample, but the calculation also works if  $X \& A$  and  $X$  are absolute numbers of people, events, or things.  $p(\text{cancer} | \text{positive})$  is a single percentage/frequency/probability, always between 0 and 1.  $(\text{positive} \& \text{cancer}) / (\text{positive})$  can be measured either in probabilities, such as 0.008/0.103, or it might be expressed in groups of women, for example 194/2494. As long as both the numerator and denominator are measured in the same units, it should make no difference.

Going from  $p(X)$  in the denominator to  $p(X \& A) + p(X \& \sim A)$  is a very straightforward step whose main purpose is as a stepping stone to the last equation. However, one common arithmetical mistake in Bayesian calculations is to divide  $p(X \& A)$  by  $p(X \& \sim A)$ , instead of dividing  $p(X \& A)$  by  $[p(X \& A) + p(X \& \sim A)]$ . For example, someone doing the breast cancer calculation tries to get the posterior probability by performing the math operation 80 / 950, instead of 80 / (80 + 950). I like to think of this as a rose-flowers error. Sometimes if you show young children a picture with eight roses and two tulips, they'll say that the picture contains more roses than flowers. (Technically, this would be called a class inclusion error.) You have to *add* the roses and the tulips to get the number of *flowers*, which you need to find the proportion of roses *within* the flowers. You can't find the proportion of roses in the tulips, or the proportion of tulips in the roses. When you look at the graph, the bottom bar consists of *all* the patients with positive results. That's what the doctor sees - a patient with a positive result. The question then becomes whether this is a healthy patient with a positive result, or a cancerous patient with a positive result. To figure the odds of that, you have to look at the proportion of cancerous patients with positive results within all patients who have positive results, because again, "a patient with a positive result" is what you actually see. You can't divide 80 by 950 because that would mean you were trying to find the proportion of cancerous patients with positive results within the group of healthy patients with positive results; it's like asking how many of the tulips are roses, instead of asking how many of the flowers are roses. Imagine using the same method to find the proportion of *healthy* patients. You would divide 950 by 80 and find that 1,187% of the patients were healthy. Or to be exact, you

would find that 1,187% of cancerous patients with positive results were healthy patients with positive results.

The last step in deriving Bayes' Theorem is going from  $p(X \& A)$  to  $p(X | A) * p(A)$ , in both the numerator and the denominator, and from  $p(X \& \sim A)$  to  $p(X | \sim A) * p(\sim A)$ , in the denominator.

Why? Well, one answer is because  $p(X|A)$ ,  $p(X|\sim A)$ , and  $p(A)$  correspond to the initial information given in all the story problems. But why were the story problems written that way?

Because in many cases,  $p(X|A)$ ,  $p(X|\sim A)$ , and  $p(A)$  are what we actually *know*; and this in turn happens because  $p(X|A)$  and  $p(X|\sim A)$  are often the quantities that directly describe *causal relations*, with the other quantities derived from them and  $p(A)$  as *statistical relations*. For example,  $p(X|A)$ , the implication from A to X, where A is what we want to know and X is our way of observing it, corresponds to the implication from a woman having breast cancer to a positive mammography. This is not just a *statistical implication* but a *direct causal relation*; a woman gets a positive mammography *because* she has breast cancer. The mammography is *designed* to detect breast cancer, and it is a fact about the physical process of the mammography exam that it has an 80% probability of detecting breast cancer. As long as the design of the mammography machine stays constant,  $p(X|A)$  will stay at 80%, even if  $p(A)$  changes - for example, if we screen a group of women with other risk factors, so that the prior frequency of women with breast cancer is 10% instead of 1%. In this case,  $p(X \& A)$  will change along with  $p(A)$ , and so will  $p(X)$ ,  $p(A|X)$ , and so on; but  $p(X|A)$  stays at 80%, because that's a fact about the mammography exam itself. (Though you do need to test this statement before relying on it; it's possible that the mammography exam might work better on some forms of breast cancer than others.)  $p(X|A)$  is one of the *simple* facts from which complex facts like  $p(X \& A)$  are constructed;  $p(X|A)$  is an *elementary* causal relation within a complex system, and it has a direct physical interpretation. This is why Bayes' Theorem has the form it does; it's not for solving math brainteasers, but for reasoning about the physical universe.

Once the derivation is finished, all the implications on the right side of the equation are of the form  $p(X|A)$  or  $p(X|\sim A)$ , while the implication on the left side is  $p(A|X)$ . As long as you remember this and you get the rest of the equation right, it shouldn't matter whether you happened to start out with  $p(A|X)$  or  $p(X|A)$  on the left side of the equation, as long as the rules are applied *consistently* - if you started out with the direction of implication  $p(X|A)$  on the left side of the equation, you would need to end up with the direction  $p(A|X)$  on the right side of the equation. This, of course, is just changing the variable labels; the point is to remember the symmetry, in order to remember the structure of Bayes' Theorem.

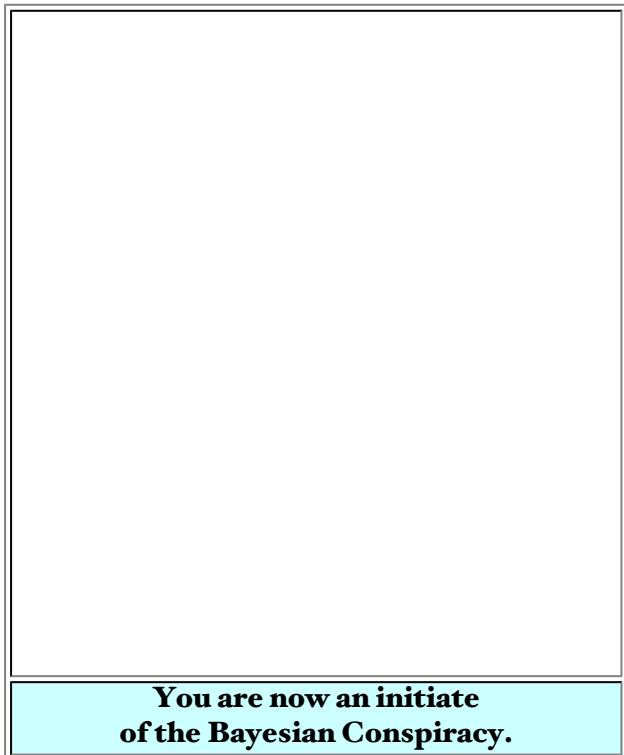
The symmetry arises because the elementary *causal relations* are generally implications from facts to observations, i.e., from breast cancer to positive mammography. The elementary *steps in reasoning* are generally implications from observations to facts, i.e., from a positive mammography to breast cancer. The left side of Bayes' Theorem is an elementary *inferential* step from the observation of positive mammography to the conclusion of an increased probability of breast cancer. Implication is written right-to-left, so we write  $p(\text{cancer}|\text{positive})$  on the left side of the equation. The right side of Bayes' Theorem describes the elementary *causal steps* - for example, from breast cancer to a positive mammography - and so the implications on the right side of Bayes' Theorem take the form  $p(\text{positive}|\text{cancer})$  or  $p(\text{positive}|\sim\text{cancer})$ .

And that's Bayes' Theorem. Rational inference on the left end, physical causality on the right end; an equation with mind on one side and reality on the other. Remember how the scientific method turned out to be a special case of Bayes' Theorem? If you wanted to put it poetically, you could say that Bayes' Theorem binds reasoning into the physical universe.

Okay, we're done.

---

**Reverend Bayes says:**



## 4. Why truth? And...<sup>1</sup>

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<sup>2</sup>.) 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.

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

1. [http://lesswrong.com/lw/go/why\\_truth\\_and/](http://lesswrong.com/lw/go/why_truth_and/)

2. <http://sl4.org/wiki/TheSimpleTruth>

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?")

## 5. What is Evidence?<sup>1</sup>

“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<sup>2</sup>. Today I’m going to talk about “evidence”. (I also intend to discuss beliefs-of-fact, not emotions or morality, as distinguished here<sup>3</sup>.)

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 “entanglement” 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

---

1. [http://lesswrong.com/lw/jl/what\\_is\\_evidence/](http://lesswrong.com/lw/jl/what_is_evidence/)

2. <http://sl4.org/wiki/TheSimpleTruth>

3. Page 625, ‘Feeling Rational’.

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.

## 6. How Much Evidence Does It Take?<sup>1</sup>

### Followup to: What is Evidence?<sup>2</sup>

Previously<sup>3</sup>, 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<sup>4</sup> 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 (0.000007%). 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<sup>5</sup> 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 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

---

1. [http://lesswrong.com/lw/jn/how\\_much\\_evidence\\_does\\_it\\_take/](http://lesswrong.com/lw/jn/how_much_evidence_does_it_take/)

2. Page 86, ‘What is Evidence?’.

3. Page 86, ‘What is Evidence?’.

4. [http://lesswrong.com/lw/bl/lotteries\\_a\\_waste\\_of\\_hope/](http://lesswrong.com/lw/bl/lotteries_a_waste_of_hope/)

5. Page 28, ‘An Intuitive Explanation of Bayes’ Theorem’.

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. Considering 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.

## 7. How to Convince Me That $2 + 2 = 3$ <sup>1</sup>

In “What is Evidence?”, I wrote<sup>2</sup>:

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<sup>3</sup>:

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 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  $XXX - XX = XX$ .

---

1. [http://lesswrong.com/lw/jr/bow\\_to\\_convince\\_me\\_that\\_2\\_2\\_3/](http://lesswrong.com/lw/jr/bow_to_convince_me_that_2_2_3/)

2. Page 86, ‘What is Evidence?’.

3. [http://lesswrong.com/lw/jl/what\\_is\\_evidence/f7b](http://lesswrong.com/lw/jl/what_is_evidence/f7b)

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<sup>4</sup>.

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.

There’s really only two possibilities, for a belief of fact<sup>5</sup>—either the belief got there via a mind-reality entangling process<sup>6</sup>, or not. If not, the belief can’t be correct except by coincidence. For beliefs with the slightest shred of internal complexity<sup>7</sup> (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.

---

4. Page 126, ‘Your Strength as a Rationalist’.

5. Page 625, ‘Feeling Rational’.

6. Page 86, ‘What is Evidence?’.

7. Page 95, ‘Occam’s Razor’.

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?

## 8. Occam's Razor<sup>1</sup>

**Followup to:** Burdensome Details<sup>2</sup>, How Much Evidence?<sup>3</sup>

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<sup>4</sup>, 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 human 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

---

1. [http://lesswrong.com/lw/jp/occams\\_razor/](http://lesswrong.com/lw/jp/occams_razor/)

2. [http://lesswrong.com/lw/jk/burdensome\\_details/](http://lesswrong.com/lw/jk/burdensome_details/)

3. Page 89, ‘How Much Evidence Does It Take?’.

4. [http://lesswrong.com/lw/jo/einstains\\_arrogance/](http://lesswrong.com/lw/jo/einstains_arrogance/)

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 “HTHHTHHTHHHTHHHTHHHT...” 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 improvement 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-

er 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!” So, even after you say “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<sup>6</sup> and anchoring<sup>7</sup> and fake explanations<sup>8</sup> and fake causality<sup>9</sup> and positive bias<sup>10</sup> and motivated cognition<sup>11</sup>, 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.

5. Page 147, ‘Fake Causality’.

6. Page 133, ‘Hindsight bias’.

7. Page 443, ‘Anchoring and Adjustment’.

8. Page 138, ‘Fake Explanations’.

9. Page 147, ‘Fake Causality’.

10. Page 164, ‘Positive Bias: Look Into the Dark’.

11. Page 505, ‘Knowing About Biases Can Hurt People’.

## 9. The Lens That Sees Its Flaws<sup>1</sup>

**Continuation of:** What is Evidence?<sup>2</sup>

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<sup>3</sup>, 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<sup>4</sup> at this, for it is indeed marvelous.

Mice see, but they don't know they have visual cortices, 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<sup>5</sup>, 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.

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

---

1. [http://lesswrong.com/lw/jm/the\\_lens\\_that\\_sees\\_its\\_flaws/](http://lesswrong.com/lw/jm/the_lens_that_sees_its_flaws/)

2. Page 86, ‘What is Evidence?’.

3. Page 154, ‘Mysterious Answers to Mysterious Questions’.

4. Page 177, “Science” as Curiosity-Stopper’.

5. <http://www.richrock.com/gifs/optical-illusion-wheels-circles-rotating.png>

separate magisterium<sup>6</sup>, far away from real life and the understanding of ordinary mortals. Science is not something that only applies to the inside of laboratories<sup>7</sup>. 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<sup>8</sup>—this thought process will not systematically correlate<sup>9</sup> optimists to branches in which no nuclear war occurs. (Some 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<sup>10</sup> 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

---

6. [http://lesswrong.com/lw/i8/religions\\_claim\\_to\\_be\\_nondisprovable/](http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/)

7. [http://lesswrong.com/lw/gv/outside\\_the\\_laboratory/](http://lesswrong.com/lw/gv/outside_the_laboratory/)

8. <http://arxiv.org/abs/astro-ph/0302131>

9. Page 86, ‘What is Evidence?’.

10. Page 28, ‘An Intuitive Explanation of Bayes’ Theorem’.

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<sup>ii</sup> 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.

---

ii. Page 625, ‘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.*



# I. Making Beliefs Pay Rent (in Anticipated Experiences)<sup>1</sup>

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 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

---

<sup>1</sup>. [http://lesswrong.com/lw/i3/making\\_beliefs\\_pay\\_rent\\_in\\_anticipated\\_experiences/](http://lesswrong.com/lw/i3/making_beliefs_pay_rent_in_anticipated_experiences/)

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 oversimplify 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 anticipation, 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<sup>1</sup>

**Followup to:** Making Beliefs Pay Rent (in Anticipated Experiences)<sup>2</sup>

Carl Sagan once told a parable<sup>3</sup> 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 more tangled than that. As discussed in yesterday’s post, there are different types of belief; not all beliefs are direct anticipations<sup>4</sup>. 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

---

1. [http://lesswrong.com/lw/i4/belief\\_in\\_belief/](http://lesswrong.com/lw/i4/belief_in_belief/)

2. Page 105, ‘Making Beliefs Pay Rent (in Anticipated Experiences)’.

3. <http://www.godlessgeeks.com/LINKS/Dragon.htm>

4. Page 105, ‘Making Beliefs Pay Rent (in Anticipated Experiences)’.

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<sup>5</sup> 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 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

---

5. Page 315, 'A Fable of Science and Politics'.

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<sup>1</sup>

You can have some fun with people whose anticipations get out of sync with what they believe<sup>2</sup>.

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.”

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 re-

---

1. [http://lesswrong.com/lw/i5/bayesian\\_judo/](http://lesswrong.com/lw/i5/bayesian_judo/)

2. Page 108, ‘Belief in Belief’.

quires.” 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<sup>1</sup>

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”<sup>2</sup> 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.

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

---

1. [http://lesswrong.com/lw/i6/professing\\_and\\_cheering/](http://lesswrong.com/lw/i6/professing_and_cheering/)

2. Page 108, ‘Belief in Belief’.

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<sup>3</sup>” isn’t a statement 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<sup>4</sup>.) 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.

---

3. Page 315, ‘A Fable of Science and Politics’.

4. Page 625, ‘Feeling Rational’.

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.

## 5. Belief as Attire<sup>1</sup>

I have so far distinguished between belief as anticipation-controller<sup>2</sup>, belief in belief<sup>3</sup>, professing and cheering<sup>4</sup>. 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<sup>5</sup> 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<sup>6</sup>, 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<sup>7</sup>. 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 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<sup>8</sup>, or religious pro-

---

1. [http://lesswrong.com/lw/iz/belief\\_as\\_attire/](http://lesswrong.com/lw/iz/belief_as_attire/)

2. Page 105, ‘Making Beliefs Pay Rent (in Anticipated Experiences)’.

3. Page 108, ‘Belief in Belief’.

4. Page 114, ‘Professing and Cheering’.

5. [http://lesswrong.com/lw/i6/professing\\_and\\_cheering/egb](http://lesswrong.com/lw/i6/professing_and_cheering/egb)

6. Page 329, ‘Are Your Enemies Innately Evil?’.

7. <http://www.imdb.com/title/tt0116629/>

8. Page 108, ‘Belief in Belief’.

fessing<sup>9</sup>, 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<sup>10</sup>. (This is the foundation on which rests the swindle of "Republicans vs. Democrats" and analogous false dilemmas<sup>11</sup> 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.

---

9. Page 114, 'Professing and Cheering'.

10. Page 315, 'A Fable of Science and Politics'.

11. Page 589, 'The Third Alternative'.

## 6. Focus Your Uncertainty<sup>1</sup>

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<sup>2</sup> 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<sup>3</sup>. 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 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 ex-*

1. [http://lesswrong.com/lw/ia/focus\\_your\\_uncertainty/](http://lesswrong.com/lw/ia/focus_your_uncertainty/)

2. [http://lesswrong.com/lw/iz/two\\_more\\_things\\_to\\_unlearn\\_from\\_school/](http://lesswrong.com/lw/iz/two_more_things_to_unlearn_from_school/)

3. [http://www.aft.org/pubs-reports/american\\_educator/issues/summer07/Crit\\_Thinking.pdf](http://www.aft.org/pubs-reports/american_educator/issues/summer07/Crit_Thinking.pdf)

*actly 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 anticipation, 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<sup>1</sup>

*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<sup>2 3</sup>

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 mis-step 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 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

---

1. [http://lesswrong.com/lw/ic/the\\_virtue\\_of\\_narrowness/](http://lesswrong.com/lw/ic/the_virtue_of_narrowness/)

2. <http://yudkowsky.net/virtues/>

3. <http://yudkowsky.net/virtues/>

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 DONT 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<sup>1</sup>

(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 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...

---

<sup>1</sup>. [http://lesswrong.com/lw/if/your\\_strength\\_as\\_a\\_rationalist/](http://lesswrong.com/lw/if/your_strength_as_a_rationalist/)

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<sup>2</sup> than a published journal article. Alas, belief is easier than disbelief; we believe instinctively, but disbelief requires a conscious effort<sup>3</sup>.

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<sup>4</sup>.

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

---

2. <http://www.overcomingbias.com/2007/08/truth-bias.html>

3. <http://www.wjh.harvard.edu/~o7Edtg/>

*Gilbert%20et%20al%20%28EVERYTHING%20YOU%20READ%29.pdf*

4. Page 105, 'Making Beliefs Pay Rent (in Anticipated Experiences)'.

## 9. Absence of Evidence Is Evidence of Absence<sup>1</sup>

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<sup>2</sup>. 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 is still higher that the *absence* of a Fifth Column would perform an absence of sabotage.

Let E stand for the observation of sabotage, H<sub>1</sub> for the hypothesis of a Japanese-American Fifth Column, and H<sub>2</sub> for the hypothesis that no Fifth Column exists. Whatever the likelihood

1. [http://lesswrong.com/lw/ib/absence\\_of\\_evidence\\_is\\_evidence\\_of\\_absence/](http://lesswrong.com/lw/ib/absence_of_evidence_is_evidence_of_absence/)

2. Page 28, 'An Intuitive Explanation of Bayes' Theorem'.

that a Fifth Column would do no sabotage, the probability  $P(E|H_1)$ , it cannot be as large as the likelihood that no Fifth Column does no sabotage, the probability  $P(E|H_2)$ . 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,  $A \rightarrow B$ , "A implies B", is not equivalent to  $\neg A \rightarrow \neg B$ , "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  $P(H| \neg E) < P(H)$ , "failure to observe E decreases the probability of H".  $P(H)$  is a weighted mix of  $P(H|E)$  and  $P(H| \neg E)$ , and necessarily lies between the two. If any of this sounds at all confusing, see An Intuitive Explanation of Bayesian Reasoning<sup>3</sup>.

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

Your strength as a rationalist<sup>4</sup> 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<sup>5</sup>. 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.

---

3. Page 28, 'An Intuitive Explanation of Bayes' Theorem'.

4. Page 126, 'Your Strength as a Rationalist'.

5. Page 108, 'Belief in Belief'.

## 10. Conservation of Expected Evidence<sup>1</sup>

**Followup to:** Absence of Evidence Is Evidence of Absence.<sup>2</sup>

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<sup>3</sup>” 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.

$$P(H) = P(H)$$

$$P(H) = P(H, E) + P(H, \neg E)$$

$$P(H) = P(H|E)*P(E) + P(H|\neg E)*P(\neg E)$$

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

---

1. [http://lesswrong.com/lw/ii/conservation\\_of\\_expected\\_evidence/](http://lesswrong.com/lw/ii/conservation_of_expected_evidence/)

2. Page 128, ‘Absence of Evidence Is Evidence of Absence’.

3. Page 128, ‘Absence of Evidence Is Evidence of Absence’.

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<sup>4</sup>.)

So if you claim<sup>5</sup> 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<sup>6</sup> 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<sup>7</sup>. 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 *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 opposite expectation of counterevidence. If you try to weaken the

---

4. Page 28, ‘An Intuitive Explanation of Bayes’ Theorem’.

5. Page 128, ‘Absence of Evidence Is Evidence of Absence’.

6. [http://lesswrong.com/lw/i8/religions\\_claim\\_to\\_be\\_nondisprovable/](http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/)

7. Page 126, ‘Your Strength as a Rationalist’.

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<sup>1</sup>

*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 (Sanchiyo 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 group was told the outcome and also 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

---

<sup>1</sup>. [http://lesswrong.com/lw/il/hindsight\\_bias/](http://lesswrong.com/lw/il/hindsight_bias/)

*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<sup>2</sup>. Or as Fischhoff (1982) put it:

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.

---

<sup>2</sup>. Page 130, ‘Conservation of Expected Evidence’.

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<sup>3</sup>. *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/apos0075/05a00120/o>

## 12. Hindsight Devalues Science<sup>1</sup>

This excerpt<sup>2</sup> 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<sup>3</sup>. (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. (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<sup>4</sup>? Take a moment to think before continuing...

1. [http://lesswrong.com/lw/im/hindsight\\_devalues\\_science/](http://lesswrong.com/lw/im/hindsight_devalues_science/)

2. [http://csm1.som.ohio-state.edu/Music829C/hindsight\\_bias.html](http://csm1.som.ohio-state.edu/Music829C/hindsight_bias.html)

3. Page 133, 'Hindsight bias'.

4. Page 130, 'Conservation of Expected Evidence'.

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<sup>5</sup> 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 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<sup>6</sup> what you *really* would have expected.

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

---

5. Page 126, ‘Your Strength as a Rationalist’.

6. Page 130, ‘Conservation of Expected Evidence’.

## 13. Fake Explanations<sup>1</sup>

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.”<sup>2</sup>

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?<sup>3</sup> The words are easily enough professed<sup>4</sup>—said in a loud, emphatic voice. But do the words actually control anticipation?<sup>5</sup>

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*.

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*.

---

1. [http://lesswrong.com/lw/ip/fake\\_explanations/](http://lesswrong.com/lw/ip/fake_explanations/)

2. Page 126, ‘Your Strength as a Rationalist’.

3. Page 117, ‘Belief as Attire’.

4. Page 114, ‘Professing and Cheering’.

5. Page 108, ‘Belief in Belief’.

And<sup>6</sup> as<sup>7</sup> we<sup>8</sup> all<sup>9</sup> know<sup>10</sup> by<sup>11</sup> this<sup>12</sup> point<sup>13</sup> (I<sup>14</sup> do<sup>15</sup> hope<sup>16</sup>), 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 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.

---

6. Page 105, ‘Making Beliefs Pay Rent (in Anticipated Experiences)’.

7. Page 108, ‘Belief in Belief’.

8. Page 112, ‘Bayesian Judo’.

9. Page 114, ‘Professing and Cheering’.

10. Page 117, ‘Belief as Attire’.

11. [http://lesswrong.com/lw/i8/religions\\_claim\\_to\\_be\\_nondisprovable/](http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/)

12. Page 119, ‘Focus Your Uncertainty’.

13. Page 126, ‘Your Strength as a Rationalist’.

14. Page 128, ‘Absence of Evidence Is Evidence of Absence’.

15. Page 136, ‘Hindsight Devalues Science’.

16. Page 130, ‘Conservation of Expected Evidence’.

## 14. Guessing the Teacher's Password<sup>1</sup>

### Followup to: Fake Explanations<sup>2</sup>

When I was young, I read popular physics books such as Richard Feynman's *QED: The Strange<sup>3</sup> 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<sup>4</sup>, 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<sup>5</sup>. 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 conduction"<sup>6</sup>) 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

1. [http://lesswrong.com/lw/iq/guessing\\_the\\_teachers\\_password/](http://lesswrong.com/lw/iq/guessing_the_teachers_password/)

2. Page 138, 'Fake Explanations'.

3. [http://lesswrong.com/lw/hs/think\\_like\\_reality/](http://lesswrong.com/lw/hs/think_like_reality/)

4. <http://www.math.utah.edu/~pa/math/polya.html>

5. [http://lesswrong.com/lw/iz/two\\_more\\_things\\_to\\_unlearn\\_from\\_school/](http://lesswrong.com/lw/iz/two_more_things_to_unlearn_from_school/)

6. Page 138, 'Fake Explanations'.

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<sup>7</sup> 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<sup>8</sup>. 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 the cooler side of the plate to try to measure where the heat goes, then this mental image of flowingness 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<sup>9</sup>, 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

---

7. Page 138, 'Fake Explanations'.

8. Page 117, 'Belief as Attire'.

9. Page 117, 'Belief as Attire'.

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?
- 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<sup>10</sup> that I am confused<sup>11</sup>. 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...

---

<sup>10</sup>. Page 126, 'Your Strength as a Rationalist'.

<sup>11</sup>. Page 136, 'Hindsight Devalues Science'.

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.*

## 15. Science as Attire<sup>1</sup>

**Prerequisites:** Fake Smallerstorm\_2

Explanations<sup>2</sup>, Belief As

Attire<sup>3</sup>

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<sup>4</sup>, 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.

I encounter people who very definitely believe in<sup>5</sup> 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 evo-

---

1. [http://lesswrong.com/lw/ir/science\\_as\\_attire/](http://lesswrong.com/lw/ir/science_as_attire/)

2. Page 138, ‘Fake Explanations’.

3. Page 117, ‘Belief as Attire’.

4. Page 126, ‘Your Strength as a Rationalist’.

5. Page 114, ‘Professing and Cheering’.

lutionary 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"<sup>6</sup> because they want to be part of the scientific in-crowd—belief as scientific attire<sup>7</sup>, 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"<sup>8</sup>. 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 cognition. 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<sup>9</sup> doesn't look to them like a lab coat; this isn't the football team they're cheering<sup>10</sup> 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

---

6. Page 138, 'Fake Explanations'.

7. Page 117, 'Belief as Attire'.

8. [http://lesswrong.com/lw/io/is\\_molecular\\_nanotechnology\\_scientific/](http://lesswrong.com/lw/io/is_molecular_nanotechnology_scientific/)

9. Page 117, 'Belief as Attire'.

10. Page 114, 'Professing and Cheering'.

true part of yourself. Anything else is probably passwords<sup>11</sup> or attire<sup>12</sup>.

---

11. Page 140, ‘Guessing the Teacher’s Password’.

12. Page 117, ‘Belief as Attire’.

## 16. Fake Causality<sup>1</sup>

**Followup to:** Fake Explanations<sup>2</sup>, Guessing the Teacher’s Password<sup>3</sup>

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<sup>4</sup>; it could explain everything.

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

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]

---

1. [http://lesswrong.com/lw/is/fake\\_causality/](http://lesswrong.com/lw/is/fake_causality/)

2. Page 138, ‘Fake Explanations’.

3. Page 140, ‘Guessing the Teacher’s Password’.

4. Page 126, ‘Your Strength as a Rationalist’.

5. Page 138, ‘Fake Explanations’.

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<sup>6</sup>, it may feel like the cause constrains<sup>7</sup> the effect, when it was merely fitted<sup>8</sup> to the effect.

Interestingly, our modern understanding of probabilistic reasoning about causality<sup>9</sup> 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 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 for-

---

6. Page 133, ‘Hindsight bias’.

7. Page 105, ‘Making Beliefs Pay Rent (in Anticipated Experiences)’.

8. Page 130, ‘Conservation of Expected Evidence’.

9. [http://books.google.com/books?id=k9VsqNz4pNYC&dq=%20pg=PP1&ots=WR9UGWdOdd&sig=w\\_Mrax-y4VVwZy5SQ-GySphNsKM&prev=http://www.google.com/search%03Fb%03Den%026safe%03Doff%026q%03Dpearl%02Bintelligent%02Bsystems%026btnG%03DSearch&sa=X&oi=print](http://books.google.com/books?id=k9VsqNz4pNYC&dq=%20pg=PP1&ots=WR9UGWdOdd&sig=w_Mrax-y4VVwZy5SQ-GySphNsKM&prev=http://www.google.com/search%03Fb%03Den%026safe%03Doff%026q%03Dpearl%02Bintelligent%02Bsystems%026btnG%03DSearch&sa=X&oi=print)

ward?” 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<sup>10</sup>“.

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.

---

<sup>10</sup>. [http://books.google.com/books?id=k9VsqNz4pNYC&dq=%26pg=PP1&ots=WR9UGWdOdd&sig=w\\_Mrax-y4VVwZy5SQ-GySphNsKMc&prev=http://www.google.com/search%3Fhl%3Den%26safe%3Doff%26q%3Dpearl%2Bintelligent%2Bsystems%26btnG%3DSearch&sa=X&oi=print](http://books.google.com/books?id=k9VsqNz4pNYC&dq=%26pg=PP1&ots=WR9UGWdOdd&sig=w_Mrax-y4VVwZy5SQ-GySphNsKMc&prev=http://www.google.com/search%3Fhl%3Den%26safe%3Doff%26q%3Dpearl%2Bintelligent%2Bsystems%26btnG%3DSearch&sa=X&oi=print)

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 book-keeping of the belief network. (Sadly, we humans can't rewrite our own code, the way a properly designed AI could.)

Speaking of "hindsight bias<sup>11</sup>" is just the nontechnical way of saying that humans do not rigorously separate forward and 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<sup>12</sup>, 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.

---

<sup>11</sup>. Page 136, 'Hindsight Devalues Science'.

<sup>12</sup>. Page 136, 'Hindsight Devalues Science'.

## 17. Semantic Stopsigns<sup>1</sup>

*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 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<sup>2</sup>, which

---

1. [http://lesswrong.com/lw/it/semantic\\_stopsigns/](http://lesswrong.com/lw/it/semantic_stopsigns/)

2. Page 114, ‘Professing and Cheering’.

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 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<sup>3</sup> 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 me, though no longer.) What distinguishes a semantic stopsign is *failure to consider the obvious next question*.

---

3. Page 625, ‘Feeling Rational’.

## 18. Mysterious Answers to Mysterious Questions<sup>1</sup>

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 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.

---

<sup>1</sup>. [http://lesswrong.com/lw/iu/mysterious\\_answers\\_to\\_mysterious\\_questions/](http://lesswrong.com/lw/iu/mysterious_answers_to_mysterious_questions/)

Calling “elan vital” an explanation, even a fake explanation<sup>2</sup> like phlogiston<sup>3</sup>, is probably giving it too much credit. It functioned primarily as a curiosity-stopper<sup>4</sup>. 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<sup>5</sup> 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<sup>6</sup> in their ignorance, made biology into a *sacred* mystery, and thereby became loath to relinquish their ignorance<sup>7</sup> 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 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

2. Page 138, ‘Fake Explanations’.

3. Page 147, ‘Fake Causality’.

4. Page 151, ‘Semantic Stopsigns’.

5. Page 147, ‘Fake Causality’.

6. [http://lesswrong.com/lw/b8/tsuyoku\\_naritai\\_i\\_want\\_to\\_become\\_stronger/](http://lesswrong.com/lw/b8/tsuyoku_naritai_i_want_to_become_stronger/)

7. <http://yudkowsky.net/virtues/>

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<sup>8</sup>—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 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<sup>9</sup> rather than an anticipation-controller<sup>10</sup>.
- 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 <sup>11</sup>this or that; but the reason why the mysterious force behaves thus is wrapped in a blank unity.

---

8. Page 119, ‘Focus Your Uncertainty’.

9. Page 151, ‘Semantic Stopsigns’.

10. Page 105, ‘Making Beliefs Pay Rent (in Anticipated Experiences)’.

11. Page 147, ‘Fake Causality’.

- Third, those who proffer the explanation cherish their ignorance<sup>12</sup>; 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.

---

12. <http://yudkowsky.net/virtues/>

## 19. The Futility of Emergence<sup>1</sup>

**Prerequisites:** Belief in Belief<sup>2</sup>, Fake Explanations<sup>3</sup>, Fake Causality<sup>4</sup>, Mysterious Answers to Mysterious Questions<sup>5</sup>

The failures of phlogiston<sup>6</sup> and vitalism<sup>7</sup> are historical<sup>8</sup> hindsight<sup>9</sup>. 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<sup>10</sup>: “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.

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

---

1. [http://lesswrong.com/lw/rv/the\\_futility\\_of\\_emergence/](http://lesswrong.com/lw/rv/the_futility_of_emergence/)

2. Page 108, ‘Belief in Belief’.

3. Page 138, ‘Fake Explanations’.

4. Page 147, ‘Fake Causality’.

5. Page 154, ‘Mysterious Answers to Mysterious Questions’.

6. Page 147, ‘Fake Causality’.

7. Page 154, ‘Mysterious Answers to Mysterious Questions’.

8. Page 136, ‘Hindsight Devalues Science’.

9. Page 133, ‘Hindsight bias’.

10. <http://en.wikipedia.org/wiki/Emergence>

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<sup>ii</sup>. 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.
- *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.

---

ii. Page 154, ‘Mysterious Answers to Mysterious Questions’.

Another fun exercise is to replace the word “emergent” with the old word<sup>12</sup>, the explanation<sup>13</sup> 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<sup>14</sup>?

“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<sup>15</sup> 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 find a way to escape the shackles<sup>16</sup> of settled science, they get up to the same shenanigans as their ancestors, dressed up in the literary genre of “science”<sup>17</sup> but still the same species psychology.

---

<sup>12</sup>. Page 140, ‘Guessing the Teacher’s Password’.

<sup>13</sup>. Page 138, ‘Fake Explanations’.

<sup>14</sup>. Page 126, ‘Your Strength as a Rationalist’.

<sup>15</sup>. Page 154, ‘Mysterious Answers to Mysterious Questions’.

<sup>16</sup>. <http://judkowsky.net/virtues/>

<sup>17</sup>. Page 117, ‘Belief as Attire’.

## 20. Say Not “Complexity”<sup>1</sup>

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 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<sup>2</sup>? ”

---

1. [http://lesswrong.com/lw/ix/say\\_not\\_complexity/](http://lesswrong.com/lw/ix/say_not_complexity/)

2. <http://yudkowsky.net/bayes/technical.html>

“Yes,” said Marcello.

“Okay,” I said, “saying ‘complexity’ doesn’t concentrate your probability mass.”

“Oh,” Marcello said, “like ‘emergence<sup>3</sup>. 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 eyelink it happens: putting a non-controlling causal node<sup>4</sup> behind something mysterious, a causal node that feels like an explanation<sup>5</sup> 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<sup>6</sup>, 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.

---

3. Page 158, ‘The Futility of Emergence’.

4. Page 147, ‘Fake Causality’.

5. Page 138, ‘Fake Explanations’.

6. Page 140, ‘Guessing the Teacher’s Password’.

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<sup>7</sup>”, 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<sup>8</sup> that you don't know how to build an AI yet, and your current life-plans will come crashing down in ruins around your ears. But perhaps I am over-explaining<sup>9</sup>, since skip-over happens by default<sup>10</sup> 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<sup>11</sup> create an illusion of understanding. Wiser to say “magic”, and leave yourself a placeholder, a reminder of work you will have to do later.

---

7. Page 158, ‘The Futility of Emergence’.

8. Page 639, ‘You Can Face Reality’.

9. Page 326, ‘Correspondence Bias’.

10. Page 140, ‘Guessing the Teacher’s Password’.

11. Page 140, ‘Guessing the Teacher’s Password’.

## 21. Positive Bias: Look Into the Dark<sup>1</sup>

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”.

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.

---

<sup>1</sup>. [http://lesswrong.com/lw/iw/positive\\_bias\\_look\\_into\\_the\\_dark/](http://lesswrong.com/lw/iw/positive_bias_look_into_the_dark/)

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<sup>2</sup> 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 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*<sup>3</sup>—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

---

2. Page 147, ‘Fake Causality’.

3. Page 126, ‘Your Strength as a Rationalist’.

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<sup>4</sup>. 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?

---

4. Page 158, ‘The Futility of Emergence’.

## 22. My Wild and Reckless Youth<sup>1</sup>

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<sup>2</sup>. 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<sup>3</sup>. 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<sup>4</sup> unto me, and possessed the same quality of wondrous impenetrability that it had at the start.

Make no mistake<sup>5</sup>, 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 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.

---

1. [http://lesswrong.com/lw/iy/my\\_wild\\_and\\_reckless\\_youth/](http://lesswrong.com/lw/iy/my_wild_and_reckless_youth/)

2. Page 154, ‘Mysterious Answers to Mysterious Questions’.

3. [http://en.wikipedia.org/wiki/Conjunction\\_fallacy](http://en.wikipedia.org/wiki/Conjunction_fallacy)

4. Page 154, ‘Mysterious Answers to Mysterious Questions’.

5. Page 326, ‘Correspondence Bias’.

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<sup>6</sup> 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<sup>7</sup> it quantum gravity. But I knew not the Way of Bayes, and did not see the level<sup>8</sup> 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 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<sup>9</sup>. You need *one*

---

6. Page 114, 'Professing and Cheering'.

7. Page 144, 'Science as Attire'.

8. Page 138, 'Fake Explanations'.

9. Page 505, 'Knowing About Biases Can Hurt People'.

*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, 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<sup>1</sup>

**Continuation of:** My Wild and Reckless Youth<sup>2</sup>

Once upon a time, in my wild and reckless youth<sup>3</sup>, when I knew not the Way of Bayes, I gave a Mysterious Answer<sup>4</sup> 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 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<sup>5</sup>, a tendency

---

1. [http://lesswrong.com/lw/iz/failing\\_to\\_learn\\_from\\_history/](http://lesswrong.com/lw/iz/failing_to_learn_from_history/)

2. Page 167, 'My Wild and Reckless Youth'.

3. Page 167, 'My Wild and Reckless Youth'.

4. Page 154, 'Mysterious Answers to Mysterious Questions'.

5. Page 326, 'Correspondence Bias'.

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.*

## 24. Making History Available<sup>1</sup>

### Followup to: Failing to Learn from History<sup>2</sup>

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<sup>3</sup>—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<sup>4</sup>.

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<sup>5</sup>. 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 you should be. Many biases involve being insufficiently moved by dry, abstract information<sup>6</sup>.

Once upon a time, I gave a Mysterious Answer<sup>7</sup> to a mysterious question, not realizing that I was making exactly the same mistake as astrologers devising mystical explanations for the stars, or al-

---

1. [http://lesswrong.com/lw/jo/making\\_history\\_available/](http://lesswrong.com/lw/jo/making_history_available/)

2. Page 170, ‘Failing to Learn from History’.

3. [http://en.wikipedia.org/wiki/Availability\\_heuristic](http://en.wikipedia.org/wiki/Availability_heuristic)

4. <http://singinst.org/Biases.pdf>

5. <http://www.overcomingbias.com/2007/07/tell-your-anti-.html>

6. [http://lesswrong.com/lw/bw/scope\\_insensitivity/](http://lesswrong.com/lw/bw/scope_insensitivity/)

7. Page 154, ‘Mysterious Answers to Mysterious Questions’.

chemists 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<sup>8</sup>, 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 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<sup>9</sup>—*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

---

8. Page 170, ‘Failing to Learn from History’.

9. [http://en.wikipedia.org/wiki/Availability\\_heuristic](http://en.wikipedia.org/wiki/Availability_heuristic)

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 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<sup>10</sup>, 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.

---

<sup>10</sup>. Page 133, 'Hindsight bias'.

## 25. Explain/Worship/Ignore?<sup>1</sup>

**Followup to:** Semantic Stopsigns<sup>2</sup>, Mysterious Answers to Mysterious Questions<sup>3</sup>

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.

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

---

1. <http://lesswrong.com/lw/jz/explainworshipignore/>

2. Page 151, ‘Semantic Stopsigns’.

3. Page 154, ‘Mysterious Answers to Mysterious Questions’.

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<sup>1</sup>

**Followup to:** Semantic Stopsigns<sup>2</sup>, Mysterious Answers to Mysterious Questions<sup>3</sup>, Say Not ‘Complexity’<sup>4</sup>

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<sup>5</sup>, so much as a curiosity-stopper<sup>6</sup>. 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<sup>7</sup>. 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?

In school, one is taught that the password<sup>8</sup> 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?

---

1. [http://lesswrong.com/lw/j3/science\\_as\\_curiositystopper/](http://lesswrong.com/lw/j3/science_as_curiositystopper/)

2. Page 151, ‘Semantic Stopsigns’.

3. Page 154, ‘Mysterious Answers to Mysterious Questions’.

4. Page 161, ‘Say Not “Complexity”’.

5. Page 138, ‘Fake Explanations’.

6. Page 151, ‘Semantic Stopsigns’.

7. Page 140, ‘Guessing the Teacher’s Password’.

8. Page 140, ‘Guessing the Teacher’s Password’.

There is, at the least, a great deal more to learn. (Physicists should ignore this paragraph and substitute a problem in evolutionary theory<sup>9</sup>, 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 de-values 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 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, “SOME-BODY ALREADY KNOWS WHY THIS ELEPHANT IS HERE<sup>10</sup>. ”

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

---

9. [http://lesswrong.com/lw/kr/an\\_alien\\_god/](http://lesswrong.com/lw/kr/an_alien_god/)

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

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?

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<sup>1</sup>

**Followup to:** Semantic Stopsigns<sup>2</sup>, We Don't Really Want Your Participation<sup>3</sup>

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?

The speaker replied:

---

1. [http://lesswrong.com/lw/jb/applause\\_lights/](http://lesswrong.com/lw/jb/applause_lights/)

2. Page 151, ‘Semantic Stopsigns’.

3. [http://lesswrong.com/lw/ja/we\\_dont\\_really\\_want\\_your\\_participation/](http://lesswrong.com/lw/ja/we_dont_really_want_your_participation/)

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<sup>4</sup> (~~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 *proposition-al* 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.

---

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

If you reverse this statement, you get:

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

Since the reversal sounds *abnormal*, 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.

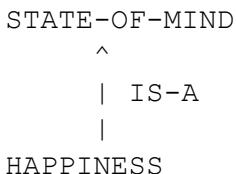
I am tempted to give a talk sometime that consists of *nothing but* applause lights, and see how long it takes for the audience to start laughing:

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<sup>1</sup>

**Followup to:** Guessing the Teacher’s Password<sup>2</sup>, Artificial Addition<sup>3</sup>

A classic paper by Drew McDermott, “Artificial Intelligence Meets Natural Stupidity<sup>4</sup>”, criticized AI programs that would try to represent notions like *happiness is a state of mind* using a semantic network:



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.*

Suppose a physicist tells you that “Light is waves<sup>5</sup>”, 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’.”

1. [http://lesswrong.com/lw/la/truly\\_part\\_of\\_you/](http://lesswrong.com/lw/la/truly_part_of_you/)

2. Page 140, ‘Guessing the Teacher’s Password’.

3. [http://lesswrong.com/lw/l9/artificial\\_addition/](http://lesswrong.com/lw/l9/artificial_addition/)

4. <http://ontolog.cim3.net/forum/ontolog-forum/2006-04/pdfb9P8wB8aYb.pdf>

5. Page 140, ‘Guessing the Teacher’s Password’.

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<sup>6</sup>?

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<sup>7</sup> 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 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 delet-

6. Page 105, ‘Making Beliefs Pay Rent (in Anticipated Experiences)’.

7. [http://lesswrong.com/lw/l9/artificial\\_addition/](http://lesswrong.com/lw/l9/artificial_addition/)

ed, 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 reprove 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 something equally novel and equally important? Would I be able to think that far outside the box<sup>8</sup>?

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<sup>9</sup> 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

---

8. Page 454, ‘The “Outside the Box” Box’.

9. <http://sl4.org/wiki/TheSimpleTruth>

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 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<sup>1</sup>

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<sup>2</sup> (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.

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

---

1. [http://lesswrong.com/lw/wb/chaotic\\_inversion/](http://lesswrong.com/lw/wb/chaotic_inversion/)

2. [http://lesswrong.com/lw/un/on\\_doing\\_the\\_impossible/](http://lesswrong.com/lw/un/on_doing_the_impossible/)

3. Page 685, 'Mind Projection Fallacy'.

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<sup>4</sup>.

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".

---

4. Page 685, 'Mind Projection Fallacy'.

## **Part III**

### **A Human's Guide to Words**

*A series on the use and abuse of words; why you often can't define a word any way you like; how human brains seem to process definitions. First introduces the Mind projection fallacy and the concept of how an algorithm feels from inside, which makes it a basic intro to key elements of the LW zeitgeist.*



## I. The Parable of the Dagger<sup>1</sup>

Once upon a time, there was a court jester who dabbled in logic.

The jester presented the king with two boxes. Upon the first box was inscribed:

“Either this box contains an angry frog, or the box with a false inscription contains an angry frog, but not both.”

On the second box was inscribed:

“Either this box contains gold and the box with a false inscription contains an angry frog, or this box contains an angry frog and the box with a true inscription contains gold.”

And the jester said to the king: “One box contains an angry frog, the other box gold; and one, and only one, of the inscriptions is true.”

The king opened the wrong box, and was savaged by an angry frog.

“You see,” the jester said, “let us hypothesize that the first inscription is the true one. Then suppose the first box contains gold. Then the other box would have an angry frog, while the box with a true inscription would contain gold, which would make the second statement true as well. Now hypothesize that the first inscription is false, and that the first box contains gold. Then the second inscription would be—”

The king ordered the jester thrown in the dungeons.

A day later, the jester was brought before the king in chains, and shown two boxes.

“One box contains a key,” said the king, “to unlock your chains; and if you find the key you are free. But the other box contains a dagger for your heart, if you fail.”

And the first box was inscribed:

---

<sup>1</sup>. [http://lesswrong.com/lw/ne/the\\_parable\\_of\\_the\\_dagger/](http://lesswrong.com/lw/ne/the_parable_of_the_dagger/)

"Either both inscriptions are true, or both inscriptions are false."

And the second box was inscribed:

"This box contains the key."

The jester reasoned thusly: "Suppose the first inscription is true. Then the second inscription must also be true. Now suppose the first inscription is false. Then again the second inscription must be true. So the second box must contain the key, if the first inscription is true, and also if the first inscription is false. Therefore, the second box must logically contain the key."

The jester opened the second box, and found a dagger.

"How?!" cried the jester in horror, as he was dragged away. "It's logically impossible!"

"It is entirely possible," replied the king. "I merely wrote those inscriptions on two boxes, and then I put the dagger in the second one."

*(Adapted from Raymond Smullyan.)*

## 2. The Parable of Hemlock<sup>1</sup>

**Followup to:** The Parable of the Dagger<sup>2</sup>

“All men are mortal. Socrates is a man. Therefore Socrates is mortal.”  
— Aristotle(?)

*Socrates raised the glass of hemlock to his lips...*

“Do you suppose,” asked one of the onlookers, “that even hemlock will not be enough to kill so wise and good a man?”

“No,” replied another bystander, a student of philosophy; “all men are mortal, and Socrates is a man; and if a mortal drink hemlock, surely he dies.”

“Well,” said the onlooker, “what if it happens that Socrates *isn’t* mortal?”

“Nonsense,” replied the student, a little sharply; “all men are mortal *by definition*; it is part of what we mean by the word ‘man’. All men are mortal, Socrates is a man, therefore Socrates is mortal. It is not merely a guess, but a *logical certainty*.”

“I suppose that’s right...” said the onlooker. “Oh, look, Socrates already drank the hemlock while we were talking.”

“Yes, he should be keeling over any minute now,” said the student.

*And they waited, and they waited, and they waited...*

“Socrates appears not to be mortal,” said the onlooker.

“Then Socrates must not be a man,” replied the student. “All men are mortal, Socrates is not mortal, therefore Socrates is not a man. And that is not merely a guess, but a *logical certainty*.”

The fundamental problem with arguing that things are true “by definition” is that you can’t make reality go a different way by choosing a different definition<sup>3</sup>.

You could reason, perhaps, as follows: “All things I have observed which wear clothing, speak language, and use tools, have also shared certain other properties as well, such as breathing air and pumping red blood. The last thirty ‘humans’ belonging to this clus-

1. [http://lesswrong.com/lw/nf/the\\_parable\\_of\\_hemlock/](http://lesswrong.com/lw/nf/the_parable_of_hemlock/)

2. Page 191, ‘The Parable of the Dagger’.

3. <http://macball.com/view.php?date=2003-04-21>

ter, whom I observed to drink hemlock, soon fell over and stopped moving. Socrates wears a toga, speaks fluent ancient Greek, and drank hemlock from a cup. So I predict that Socrates will keel over in the next five minutes.”

But that would be mere *guessing*. It wouldn’t be, y’know, absolutely and eternally certain<sup>4</sup>. The Greek philosophers—like most prescientific philosophers—were rather fond of certainty.

Luckily the Greek philosophers have a crushing rejoinder to your questioning. You have misunderstood the meaning of “All humans are mortal,” they say. It is not a mere *observation*. It is part of the *definition* of the word “human”. Mortality is one of several properties that are individually necessary, and together sufficient, to determine membership in the class “human”. The statement “All humans are mortal” is a logically valid truth, absolutely unquestionable. And if Socrates is human, he *must* be mortal: it is a logical deduction, as certain as certain can be.

But then we can never know for certain that Socrates is a “human” until after Socrates has been observed to be mortal. It does no good to observe that Socrates speaks fluent Greek, or that Socrates has red blood, or even that Socrates has human DNA. None of these characteristics are *logically equivalent* to mortality. You have to *see him die* before you can conclude that he was human.

(And even then it’s not infinitely certain<sup>5</sup>. What if Socrates rises from the grave a night after you see him die? Or more realistically, what if Socrates is signed up for cryonics? If mortality is defined to mean finite lifespan, then you can never really *know* if someone was human, until you’ve observed to the end of eternity—just to make sure they don’t come back. Or you could *think* you saw Socrates keel over, but it could be an illusion projected onto your eyes with a retinal scanner. Or maybe you just hallucinated the whole thing...)

The problem with syllogisms is that they’re *always* valid. “All humans are mortal; Socrates is human; therefore Socrates is mortal” is—if you treat it as a logical syllogism—logically valid within our own universe. It’s also logically valid within neighboring Everett branches in which, due to a slightly different evolved biochemistry,

---

4. Page 604, ‘Absolute Authority’.

5. Page 614, ‘Infinite Certainty’.

hemlock is a delicious treat rather than a poison. And it's logically valid even in universes where Socrates never existed, or for that matter, where humans never existed.

The Bayesian definition<sup>6</sup> of evidence favoring a hypothesis is evidence which we are more likely to see if the hypothesis is true than if it is false. Observing that a syllogism is logically valid can never be evidence favoring any empirical proposition, because the syllogism will be logically valid whether that proposition is true or false.

Syllogisms are valid in all possible worlds, and therefore, observing their validity never tells us anything about *which* possible world we actually live in.

This doesn't mean that logic is useless—just that logic can only tell us that which, *in some sense*, we already know. But we do not always believe what we know. Is the number 29384209 prime? By virtue of how I define my decimal system and my axioms of arithmetic, I have already determined my answer to this question—but I do not know what my answer is yet, and I must do some logic to find out.

Similarly, if I form the uncertain empirical generalization “Humans are vulnerable to hemlock”, and the uncertain empirical guess “Socrates is human”, logic can tell me that my previous guesses are predicting that Socrates will be vulnerable to hemlock.

It's been suggested that we can view logical reasoning as resolving our uncertainty about impossible possible worlds—eliminating probability mass in logically impossible worlds which we did not know to be logically impossible. In this sense, logical argument can be treated as observation<sup>7</sup>.

But when you talk about an empirical prediction like “Socrates is going to keel over and stop breathing” or “Socrates is going to do fifty jumping jacks and then compete in the Olympics next year”, that is a matter of possible worlds, not impossible possible worlds.

Logic can tell us which hypotheses match up to which observations, and it can tell us what these hypotheses predict for the future—it can bring old observations and previous guesses to bear on a new problem. But logic never flatly says, “Socrates *will* stop breathing now.” Logic never dictates any empirical question; it nev-

---

6. Page 28, ‘An Intuitive Explanation of Bayes’ Theorem’.

7. Page 92, ‘How to Convince Me That  $2 + 2 = 3$ ’.

er settles any real-world query which could, by any stretch of the imagination, go either way.

Just remember the Litany Against Logic:

Logic stays true, wherever you may go,  
So logic never tells you where you live.

### 3. Words as Hidden Inferences<sup>1</sup>

#### Followup to: The Parable of Hemlock<sup>2</sup>

Suppose I find a barrel, sealed at the top, but with a hole large enough for a hand. I reach in, and feel a small, curved object. I pull the object out, and it's blue—a bluish egg. Next I reach in and feel something hard and flat, with edges—which, when I extract it, proves to be a red cube. I pull out 11 eggs and 8 cubes, and every egg is blue, and every cube is red.

Now I reach in and I feel another egg-shaped object. Before I pull it out and look, I have to guess: What will it look like?

The evidence doesn't prove that every egg in the barrel is blue, and every cube is red. The evidence doesn't even argue this all that strongly: 19 is not a large sample size. Nonetheless, I'll guess that this egg-shaped object is blue—or as a runner-up guess, red. If I guess anything else, there's as many possibilities as distinguishable colors—and for that matter, who says the egg has to be a single shade? Maybe it has a picture of a horse painted on.

So I say “blue”, with a dutiful patina of humility. For I am a sophisticated rationalist-type person, and I keep track of my assumptions and dependencies—I guess, but I'm aware that I'm guessing... right?

But when a large yellow striped feline-shaped object leaps out at me from the shadows, I think, “Yikes! A tiger!” Not, “Hm... objects with the properties of largeness, yellowness, stripedness, and feline shape, have previously often possessed the properties ‘hungry’ and ‘dangerous’, and thus, although it is not logically necessary, it may be an empirically good guess that *aaauuughhhh CRUNCH CRUNCH GULP.*”

The human brain, for some odd reason, seems to have been adapted to make this inference quickly, automatically, and without keeping explicit track of its assumptions.

And if I name the egg-shaped objects “bleggs” (for blue eggs) and the red cubes “rubes”, then, when I reach in and feel another egg-shaped object, I may think: *Oh, it's a blegg*, rather than considering all that problem-of-induction stuff.

---

1. [http://lesswrong.com/lw/ng/words\\_as\\_hidden\\_inferences/](http://lesswrong.com/lw/ng/words_as_hidden_inferences/)

2. Page 193, ‘The Parable of Hemlock’.

It is a common misconception that you can define a word any way you like.

This would be true *if* the brain treated words as purely logical constructs, Aristotelian classes, and you never took out any more information than you put in<sup>3</sup>.

Yet the brain goes on about its work of categorization, whether or not we consciously approve. “All humans are mortal, Socrates is a human, therefore Socrates is mortal”—thus spake the ancient Greek philosophers. Well, if mortality is part of your logical definition of “human”, you can’t logically classify Socrates as human until you observe him to be mortal<sup>4</sup>. But—this is the problem—Aristotle knew perfectly well that Socrates was a human. Aristotle’s brain placed Socrates in the “human” category as efficiently as your own brain categorizes tigers, apples, and everything else in its environment: Swiftly, silently, and without conscious approval.

Aristotle laid down rules under which no one could conclude Socrates was “human” until after he died. Nonetheless, Aristotle and his students went on concluding that living people were humans and therefore mortal; they saw distinguishing properties such as human faces and human bodies, and their brains made the leap to inferred properties such as mortality.

Misunderstanding the working of your own mind does *not*, thankfully, prevent the mind from doing its work. Otherwise Aristotelians would have starved, unable to conclude that an object was edible merely because it looked and felt like a banana.

So the Aristotelians went on classifying environmental objects on the basis of partial information, the way people had always done. Students of Aristotelian logic went on thinking exactly the same way, but they had acquired an erroneous picture of *what* they were doing.

If you asked an Aristotelian philosopher whether Carol the grocer was mortal, they would say “Yes.” If you asked them how they knew, they would say “All humans are mortal, Carol is human, therefore Carol is mortal.” Ask them whether it was a guess or a certainty, and they would say it was a certainty (if you asked before

---

3. Page 193, ‘The Parable of Hemlock’.

4. Page 193, ‘The Parable of Hemlock’.

the sixteenth century, at least). Ask them how they knew that humans were mortal, and they would say it was established by definition.

The Aristotelians were still the same people, they retained their original natures, but they had acquired incorrect beliefs about their own functioning<sup>5</sup>. They looked into the mirror of self-awareness, and saw something unlike their true selves: they reflected incorrectly.

Your brain doesn't treat words as logical definitions with no empirical consequences, and so neither should you. The mere act of creating a word can cause your mind to allocate a category, and thereby trigger unconscious inferences of similarity. Or block inferences of similarity; if I create two labels<sup>6</sup> I can get your mind to allocate two categories. Notice how I said "you" and "your brain" as if they were different things?

Making errors about the inside of your head doesn't change what's there; otherwise Aristotle would have died when he concluded that the brain was an organ for cooling the blood. Philosophical mistakes usually don't interfere with blink-of-an-eye perceptual inferences.

But philosophical mistakes can severely mess up the deliberate thinking processes that we use to try to correct our first impressions. If you believe that you can "define a word any way you like", without realizing that your brain goes on categorizing without your conscious oversight, then you won't take the effort to choose your definitions wisely.

---

5. Page 99, 'The Lens That Sees Its Flaws'.

6. [http://lesswrong.com/lw/mg/the\\_twoparty\\_swindle/](http://lesswrong.com/lw/mg/the_twoparty_swindle/)

## 4. Extensions and Intensions<sup>1</sup>

**Followup to:** Words as Hidden Inferences<sup>2</sup>

“What is red?”  
“Red is a color.”  
“What’s a color?”  
“A color is a property of a thing.”

But what is a thing? And what’s a property? Soon the two are lost in a maze of words defined in other words, the problem that Steven Harnad once described<sup>3</sup> as trying to learn Chinese from a Chinese/Chinese dictionary.

Alternatively, if you asked me “What is red?” I could point to a stop sign, then to someone wearing a red shirt, and a traffic light that happens to be red, and blood from where I accidentally cut myself, and a red business card, and then I could call up a color wheel on my computer and move the cursor to the red area. This would probably be sufficient, though if you know what the word “No” means, the truly strict<sup>4</sup> would insist that I point to the sky and say “No.”

I think I stole this example from S. I. Hayakawa—though I’m really not sure, because I heard this way back in the indistinct blur of my childhood. (When I was 12, my father accidentally deleted all my computer files. I have no memory of anything before that.)

But that’s how I remember first learning about the difference between intensional and extensional definition. To give an “intensional definition” is to define a word or phrase in terms of other words, as a dictionary does. To give an “extensional definition” is to point to examples, as adults do when teaching children. The preceding sentence gives an intensional definition of “extensional definition”, which makes it an extensional example of “intensional definition”.

---

1. [http://lesswrong.com/lw/nb/extensions\\_and\\_intensions/](http://lesswrong.com/lw/nb/extensions_and_intensions/)

2. Page 197, ‘Words as Hidden Inferences’.

3. <http://www.ecs.soton.ac.uk/%07Ebarnad/Papers/Harnad/barnad90.sgproblem.html>

4. Page 164, ‘Positive Bias: Look Into the Dark’.

In Hollywood Rationality and popular culture generally, “rationalists” are depicted as word-obsessed, floating in endless verbal space disconnected from reality.

But the actual Traditional Rationalists have long insisted on maintaining a tight connection to experience:

“If you look into a textbook of chemistry for a definition of lithium, you may be told that it is that element whose atomic weight is 7 very nearly. But if the author has a more logical mind he will tell you that if you search among minerals that are vitreous, translucent, grey or white, very hard, brittle, and insoluble, for one which imparts a crimson tinge to an unluminous flame, this mineral being triturated with lime or witherite rats-bane, and then fused, can be partly dissolved in muriatic acid; and if this solution be evaporated, and the residue be extracted with sulphuric acid, and duly purified, it can be converted by ordinary methods into a chloride, which being obtained in the solid state, fused, and electrolyzed with half a dozen powerful cells, will yield a globule of a pinkish silvery metal that will float on gasolene; and the material of that is a specimen of lithium.”

— Charles Sanders Peirce

That’s an example of “logical mind” as described by a genuine Traditional Rationalist, rather than a Hollywood scriptwriter.

But note: Peirce isn’t *actually* showing you a piece of lithium. He didn’t have pieces of lithium stapled to his book. Rather he’s giving you a treasure map—an intensionally defined procedure which, when executed, will lead you to an extensional example of lithium. This is not the same as just tossing you a hunk of lithium, but it’s not the same as saying “atomic weight 7” either. (Though if you had *sufficiently sharp* eyes, saying “3 protons” might let you pick out lithium at a glance...)

So that is intensional and extensional *definition*, which is a way of telling someone else what you mean by a concept. When I talked about “definitions” above, I talked about a way of *communicating* concepts—*telling someone else* what you mean by “red”, “tiger”,

“human”, or “lithium”. Now let’s talk about the actual concepts themselves.

The actual intension of my “tiger” concept would be the neural pattern (in my temporal cortex) that inspects an incoming signal from the visual cortex to determine whether or not it is a tiger.

The actual extension of my “tiger” concept is everything I call a tiger.

Intensional definitions don’t capture entire intensions; extensional definitions don’t capture entire extensions. If I point to just one tiger and say the word “tiger”, the communication may fail if they think I mean “dangerous animal” or “male tiger” or “yellow thing”. Similarly, if I say “dangerous yellow-black striped animal”, without pointing to anything, the listener may visualize giant hornets.

You can’t capture in words all the details of the cognitive concept—as it exists in your mind—that lets you recognize things as tigers or nontigers. It’s too large. And you can’t point to all the tigers you’ve ever seen, let alone everything you *would* call a tiger.

The strongest definitions use a crossfire of intensional and extensional communication to nail down a concept. Even so, you only communicate *maps to* concepts, or instructions for building concepts—you don’t communicate the *actual* categories as they exist in your mind or in the world.

(Yes, with enough creativity you can construct exceptions to this rule, like “Sentences Eliezer Yudkowsky has published containing the term ‘*huragaloni*’ as of Feb 4, 2008”. I’ve just shown you this concept’s entire extension. But except in mathematics, definitions are usually treasure maps, not treasure.)

So that’s another reason you can’t “define a word any way you like”: You can’t directly program concepts into someone else’s brain.

Even within the Aristotelian paradigm, where we pretend that the definitions are the actual concepts, you don’t have *simultaneous* freedom of intension and extension. Suppose I define Mars as “A huge red rocky sphere, around a tenth of Earth’s mass and 50% further away from the Sun”. It’s then a separate matter to show that this intensional definition matches some particular extensional thing in my experience, or indeed, that it matches any real thing

whatsoever. If instead I say “That’s Mars” and point to a red light in the night sky, it becomes a separate matter to show that this extensional light matches any particular intensional definition I may propose—or any intensional beliefs I may have—such as “Mars is the God of War”.

But most of the brain’s work of applying intensions happens sub-deliberately. We aren’t consciously aware that our identification of a red light as “Mars” is a separate matter from our verbal definition “Mars is the God of War”. No matter what kind of intensional definition I make up to describe Mars, my mind believes that “Mars” refers to this thingy<sup>5</sup>, and that it is the fourth planet in the Solar System.

When you take into account the way the human mind actually, pragmatically works, the notion “I can define a word any way I like” soon becomes “I can believe anything I want about a fixed set of objects” or “I can move any object I want in or out of a fixed membership test”. Just as you can’t usually convey a concept’s whole intension in words because it’s a big complicated neural membership test, you can’t *control* the concept’s entire intension because it’s applied sub-deliberately. This is why arguing that XYZ is true “by definition” is so popular. If definition changes behaved like the empirical nullops they’re supposed to be, no one would bother arguing them. But abuse definitions just a little, and they turn into magic wands—in arguments, of course; not in reality.

---

5. [http://en.wikipedia.org/wiki/Image:Mars\\_Hubble.jpg](http://en.wikipedia.org/wiki/Image:Mars_Hubble.jpg)

## 5. Similarity Clusters<sup>1</sup>

### Followup to: Extensions and Intensions<sup>2</sup>

Once upon a time, the philosophers of Plato's Academy claimed that the best definition of human was a "featherless biped". Diogenes of Sinope, also called Diogenes the Cynic, is said to have promptly exhibited a plucked chicken and declared "Here is Plato's man." The Platonists promptly changed their definition to "a featherless biped with broad nails".

No dictionary, no encyclopedia, has ever listed all the things that humans have in common. We have red blood, five fingers on each of two hands, bony skulls, 23 pairs of chromosomes—but the same might be said of other animal species. We make complex tools to make complex tools, we use syntactical combinatorial language, we harness critical fission reactions as a source of energy: these things may serve out to single out only humans, but not all humans—many of us have never built a fission reactor. With the right set of necessary-and-sufficient gene sequences you could single out all humans, and only humans—at least for now—but it would still be far from *all* that humans have in common.

But so long as you don't happen to be near a plucked chicken, saying "Look for featherless bipeds" may serve to pick out a few dozen of the particular things that are humans, as opposed to houses, vases, sandwiches, cats, colors, or mathematical theorems.

Once the definition "featherless biped" has been bound to some *particular* featherless bipeds, you can look over the group, and begin harvesting some of the *other* characteristics—beyond mere feather-free twolegginess—that the "featherless bipeds" seem to share in common. The particular featherless bipeds that you see seem to also use language, build complex tools, speak combinatorial language with syntax, bleed red blood if poked, die when they drink hemlock.

Thus the category "human" grows richer, and adds more and more characteristics; and when Diogenes finally presents his plucked chicken, we are not fooled: This plucked chicken is obviously not similar to the other "featherless bipeds".

---

1. [http://lesswrong.com/lw/nj/similarity\\_clusters/](http://lesswrong.com/lw/nj/similarity_clusters/)

2. Page 200, 'Extensions and Intensions'.

(If Aristotelian logic were a good model of human psychology, the Platonists would have looked at the plucked chicken and said, “Yes, that’s a human; what’s your point?”)

If the *first* featherless biped you see is a plucked chicken, then you may end up thinking that the verbal label “human” denotes a plucked chicken; so I can modify my treasure map to point to “featherless bipeds with broad nails”, and if I am wise, go on to say, “See Diogenes over there? That’s a human, and I’m a human, and you’re a human; and that chimpanzee is not a human, though fairly close.”

The initial clue only has to lead the user to the similarity cluster—the group of things that have many characteristics in common. After that, the initial clue has served its purpose, and I can go on to convey the new information “humans are currently mortal”, or whatever else I want to say about us featherless bipeds.

A dictionary is best thought of, not as a book of Aristotelian class definitions, but a book of hints for matching verbal labels to similarity clusters, or matching labels to properties that are useful in distinguishing similarity clusters.

## 6. Typicality and Asymmetrical Similarity<sup>1</sup>

### Followup to: Similarity Clusters<sup>2</sup>

Birds fly. Well, except ostriches don't. But which is a more typical bird—a robin, or an ostrich?

Which is a more typical chair: A desk chair, a rocking chair, or a beanbag chair?

Most people would say that a robin is a more typical bird, and a desk chair is a more typical chair. The cognitive psychologists who study this sort of thing experimentally, do so under the heading of “typicality effects” or “prototype effects” (Rosch and Lloyd 1978). For example, if you ask subjects to press a button to indicate “true” or “false” in response to statements like “A robin is a bird” or “A penguin is a bird”, reaction times are faster for more central examples. (I’m still unpacking my books, but I’m reasonably sure my source on this is Lakoff 1986.) Typicality measures correlate well using different investigative methods—reaction times are one example; you can also ask people to directly rate, on a scale of 1 to 10, how well an example (like a specific robin) fits a category (like “bird”).

So we have a mental measure of typicality—which might, perhaps, function as a heuristic—but is there a corresponding bias we can use to pin it down?

Well, which of these statements strikes you as more natural: “98 is approximately 100”, or “100 is approximately 98”? If you’re like most people, the first statement seems to make more sense. (Sadock 1977.) For similar reasons, people asked to rate how similar Mexico is to the United States, gave consistently higher ratings than people asked to rate how similar the United States is to Mexico. (Tversky and Gati 1978.)

And if that still seems harmless, a study by Rips (1975) showed that people were more likely to expect a disease would spread from robins to ducks on an island, than from ducks to robins. Now this is not a *logical* impossibility, but in a pragmatic sense, whatever difference separates a duck from a robin and would make a disease less likely to spread from a duck to a robin, must also be a difference

---

1. [http://lesswrong.com/lw/nk/typicality\\_and\\_asymmetrical\\_similarity/](http://lesswrong.com/lw/nk/typicality_and_asymmetrical_similarity/)

2. Page 204, ‘Similarity Clusters’.

between a robin and a duck, and would make a disease less likely to spread from a robin to a duck.

Yes, you can come up with rationalizations, like “Well, there could be more neighboring species of the robins, which would make the disease more likely to spread initially, etc.,” but be careful not to try too hard to rationalize the probability ratings of subjects who didn’t even realize there was a comparison going on. And don’t forget that Mexico is more similar to the United States than the United States is to Mexico, and that 98 is closer to 100 than 100 is to 98. A simpler interpretation is that people are using the (demonstrated) similarity heuristic as a proxy for the probability that a disease spreads, and this heuristic is (demonstrably) asymmetrical.

Kansas is unusually close to the center of the United States, and Alaska is unusually far from the center of the United States; so Kansas is probably closer to most places in the US and Alaska is probably farther. It does not follow, however, that Kansas is closer to Alaska than is Alaska to Kansas. But people seem to reason (metaphorically speaking) as if closeness is an inherent property of Kansas and distance is an inherent property of Alaska; so that Kansas is still close, even to Alaska; and Alaska is still distant, even from Kansas.

So once again we see that Aristotle’s notion of categories—logical classes with membership determined by a collection of properties that are individually strictly necessary, and together strictly sufficient—is not a good model of human cognitive psychology. (Science’s view has changed somewhat over the last 2350 years? Who would’ve thought?) We don’t even reason as if set membership is a true-or-false property: Statements of set membership can be more or less true. (Note: This is *not* the same thing as being more or less probable.)

One more reason not to pretend<sup>3</sup> that you, or anyone else, is *really* going to treat words as Aristotelian logical classes.

---

Lakoff, George. (1986). *Women, Fire and Dangerous Things: What Categories Tell Us About the Nature of Thought*. University of Chicago Press, Chicago.

3. Page 197, ‘Words as Hidden Inferences’.

- Rips, Lance J. (1975). "Inductive judgments about natural categories." *Journal of Verbal Learning and Verbal Behavior*. 14:665-81.
- Rosch, Eleanor and B. B. Lloyd, eds. (1978). *Cognition and Categorization*. Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Sadock, Jerrold. (1977). "Truth and Approximations." In *Papers from the Third Annual Meeting of the Berkeley Linguistics Society*, pp. 430-39. Berkeley: Berkeley Linguistics Society.
- Tversky, Amos and Itamar Gati. (1978). "Studies of Similarity". In Rosch and Lloyd (1978).

## 7. The Cluster Structure of Thingspace<sup>1</sup>

**Followup to:** Typicality and Asymmetrical Similarity<sup>2</sup>

The notion of a “configuration space” is a way of translating object *descriptions* into object *positions*. It may seem like **blue** is “closer” to **blue-green** than to **red**, but how much closer? It’s hard to answer that question by just staring at the colors. But it helps to know that the (proportional) color coordinates in RGB are  $0:0:5$ ,  $0:3:2$  and  $5:0:0$ . It would be even clearer if plotted on a 3D graph.

In the same way, you can see a robin as a robin—brown tail, red breast, standard robin shape, maximum flying speed when unladen, its species-typical DNA and individual alleles. Or you could see a robin as a single point in a configuration space whose dimensions described everything we knew, or could know, about the robin.

A robin is bigger than a virus, and smaller than an aircraft carrier—that might be the “volume” dimension. Likewise a robin weighs more than a hydrogen atom, and less than a galaxy; that might be the “mass” dimension. Different robins will have strong correlations between “volume” and “mass”, so the robin-points will be lined up in a fairly linear string, in those two dimensions—but the correlation won’t be exact, so we do need two separate dimensions.

This is the benefit of viewing robins as points in space: You couldn’t see the linear lineup as easily if you were just imagining the robins as cute little wing-flapping creatures.

A robin’s DNA is a highly multidimensional variable, but you can still think of it as part of a robin’s location in thingspace—millions of quaternary coordinates, one coordinate for each DNA base—or maybe a more sophisticated view that . The shape of the robin, and its color (surface reflectance), you can likewise think of as part of the robin’s position in thingspace, even though they aren’t *single* dimensions.

Just like the coordinate point  $0:0:5$  contains the same information as the actual HTML color **blue**, we shouldn’t actually lose information when we see robins as points in space. We believe the same statement about the robin’s mass whether we visualize a robin

---

1. [http://lesswrong.com/lw/nl/the\\_cluster\\_structure\\_of\\_thingspace/](http://lesswrong.com/lw/nl/the_cluster_structure_of_thingspace/)

2. Page 206, ‘Typicality and Asymmetrical Similarity’.

balancing the scales opposite a 0.07-kilogram weight, or a robin-point with a mass-coordinate of +70.

We can even imagine a configuration space with one or more dimensions for every distinct characteristic of an object, so that the *position* of an object's point in this space corresponds to *all* the information in the real object itself. Rather redundantly represented, too—dimensions would include the mass, the volume, and the density.

If you think that's extravagant, quantum physicists use an *infinite-dimensional* configuration space, and a single point in that space describes the location of every particle in the universe. So we're actually being comparatively conservative in our visualization of *thingspace*—a point in thingspace describes just one object, not the entire universe.

If we're not sure of the robin's exact mass and volume, then we can think of a little cloud in thingspace, a *volume of uncertainty*, within which the robin might be. The density of the cloud is the density of our belief that the robin has that particular mass and volume. If you're more sure of the robin's density than of its mass and volume, your probability-cloud will be highly concentrated in the density dimension, and concentrated around a slanting line in the subspace of mass/volume. (Indeed, the cloud here is actually a surface, because of the relation  $VD = M$ .)

"Radial categories" are how cognitive psychologists describe the non-Aristotelian boundaries of words. The central "mother" conceives her child, gives birth to it, and supports it. Is an egg donor who never sees her child a mother? She is the "genetic mother". What about a woman who is implanted with a foreign embryo and bears it to term? She is a "surrogate mother". And the woman who raises a child that isn't hers genetically? Why, she's an "adoptive mother". The Aristotelian syllogism would run, "Humans have ten fingers, Fred has nine fingers, therefore Fred is not a human" but the way we actually think is "Humans have ten fingers, Fred is a human, therefore Fred is a 'nine-fingered human'."

We can think about the radial-ness of categories in intensional terms, as described above—properties that are usually present, but optionally absent. If we thought about the intension of the word "mother", it might be like a distributed glow in thingspace, a glow

whose intensity matches the degree to which that volume of thingspace matches the category “mother”. The glow is concentrated in the center of genetics and birth and child-raising; the volume of egg donors would also glow, but less brightly.

Or we can think about the radial-ness of categories extensionally. Suppose we mapped all the birds in the world into thingspace, using a distance metric that corresponds as well as possible to perceived similarity in humans: A robin is more similar to another robin, than either is similar to a pigeon, but robins and pigeons are all more similar to each other than either is to a penguin, etcetera.

Then the center of all birdness would be densely populated by many neighboring tight clusters, robins and sparrows and canaries and pigeons and many other species. Eagles and falcons and other large predatory birds would occupy a nearby cluster. Penguins would be in a more distant cluster, and likewise chickens and ostriches.

The result might look, indeed, something like an astronomical cluster: many galaxies orbiting the center, and a few outliers.

Or we could think simultaneously about both the intension of the cognitive category “bird”, and its extension in real-world birds: The central clusters of robins and sparrows glowing brightly with highly typical birdness; satellite clusters of ostriches and penguins glowing more dimly with atypical birdness, and Abraham Lincoln a few megaparsecs away and glowing not at all.

I prefer that last visualization—the glowing points—because as I see it, the structure of the cognitive intension followed from the extensional cluster structure. First came the structure-in-the-world, the empirical distribution of birds over thingspace; then, by observing it, we formed a category whose intensional glow roughly overlays this structure.

This gives us yet another view of why words are not Aristotelian classes: the empirical clustered structure of the real universe is not so crystalline. A natural cluster, a group of things highly similar to each other, may have *no* set of necessary and sufficient properties—*no* set of characteristics that all group members have, and no non-members have.

But even if a category is irrecoverably blurry and bumpy, there's no need to panic. I would not object if someone said that birds

are “feathered flying things”. *But penguins don’t fly!*—well, fine. The usual rule has an exception; it’s not the end of the world. Definitions can’t be expected to exactly match the empirical structure of thingspace in any event, because the map is smaller and much less complicated than the territory. The point of the definition “feathered flying things” is to lead the listener to the bird cluster, not to give a total description of every existing bird down to the molecular level.

When you draw a boundary around a group of extensional points *empirically* clustered in thingspace, you may find at least one exception to every simple intensional rule you can invent.

But if a definition works well enough in practice to point out the intended empirical cluster, objecting to it may justly be called “nitpicking”.

## 8. Disguised Queries<sup>1</sup>

**Followup to:** The Cluster Structure of Thingspace<sup>2</sup>

Imagine that you have a peculiar job in a peculiar factory: Your task is to take objects from a mysterious conveyor belt, and sort the objects into two bins. When you first arrive, Susan the Senior Sorter explains to you that blue egg-shaped objects are called “bleggs” and go in the “blegg bin”, while red cubes are called “rubes” and go in the “rube bin”.

Once you start working, you notice that bleggs and rubes differ in ways besides color and shape. Bleggs have fur on their surface, while rubes are smooth. Bleggs flex slightly to the touch; rubes are hard. Bleggs are opaque; the rube’s surface slightly translucent.

Soon after you begin working, you encounter a blegg shaded an unusually dark blue—in fact, on closer examination, the color proves to be purple, halfway between red and blue.

Yet wait! Why are you calling this object a “blegg”? A “blegg” was originally defined as blue and egg-shaped—the qualification of blueness appears in the very name “blegg”, in fact. This object is not blue. One of the necessary qualifications is missing; you should call this a “purple egg-shaped object”, not a “blegg”.

But it so happens that, in addition to being purple and egg-shaped, the object is also furred, flexible, and opaque. So when you saw the object, you thought, “Oh, a strangely colored blegg.” It certainly isn’t a rube... right?

Still, you aren’t quite sure what to do next. So you call over Susan the Senior Sorter.

“Oh, yes, it’s a blegg,” Susan says, “you can put it in the blegg bin.”

You start to toss the purple blegg into the blegg bin, but pause for a moment. “Susan,” you say, “how do you know this is a blegg?”

Susan looks at you oddly. “Isn’t it obvious? This object may be purple, but it’s still egg-shaped, furred, flexible, and opaque, like all the other bleggs. You’ve got

---

1. [http://lesswrong.com/lw/nm/disguised\\_queries/](http://lesswrong.com/lw/nm/disguised_queries/)

2. Page 209, ‘The Cluster Structure of Thingspace’.

to expect a few color defects. Or is this one of those philosophical conundrums, like 'How do you know the world wasn't created five minutes ago complete with false memories?' In a philosophical sense I'm not *absolutely certain* that this is a blegg, but it seems like a good guess."

"No, I mean..." You pause, searching for words. "Why is there a blegg bin and a rube bin? What's the *difference* between bleggs and rubes?"

"Bleggs are blue and egg-shaped, rubes are red and cube-shaped," Susan says patiently. "You got the standard orientation lecture, right?"

"Why do bleggs and rubes *need* to be sorted?"

"Er... because otherwise they'd be all mixed up?" says Susan. "Because nobody will pay us to sit around all day and *not* sort bleggs and rubes?"

"Who originally determined that the first blue egg-shaped object was a 'blegg', and how did they determine that?"

Susan shrugs. "I suppose you could just as easily call the red cube-shaped objects 'bleggs' and the blue egg-shaped objects 'rubes', but it seems easier to remember this way."

You think for a moment. "Suppose a completely mixed-up object came off the conveyor. Like, an orange sphere-shaped furred translucent object with writhing green tentacles. How could I tell whether it was a blegg or a rube?"

"Wow, no one's ever found an object *that* mixed up," says Susan, "but I guess we'd take it to the sorting scanner."

"How does the sorting scanner work?" you inquire. "X-rays? Magnetic resonance imaging? Fast neutron transmission spectroscopy?"

"I'm told it works by Bayes's Rule, but I don't quite understand how," says Susan. "I like to say it, though. Bayes Bayes Bayes Bayes Bayes."

"What does the sorting scanner *tell* you?"

"It tells you whether to put the object into the blegg

bin or the rube bin. That's why it's called a sorting scanner."

At this point you fall silent.

"Incidentally," Susan says casually, "it may interest you to know that bleggs contain small nuggets of vanadium ore, and rubes contain shreds of palladium, both of which are useful industrially."

"Susan, you are pure evil."

"Thank you."

So now it seems we've discovered the heart and essence of bleggness: a blegg is an object that contains a nugget of vanadium ore. Surface characteristics, like blue color and furredness, do not *determine* whether an object is a blegg; surface characteristics only matter because they help you *infer* whether an object is a blegg, that is, whether the object contains vanadium.

Containing vanadium is a necessary and sufficient definition: all bleggs contain vanadium and everything that contains vanadium is a blegg: "blegg" is just a shorthand way of saying "vanadium-containing object." Right?

Not so fast, says Susan: Around 98% of bleggs contain vanadium, but 2% contain palladium instead. To be precise (Susan continues) around 98% of blue egg-shaped furred flexible opaque objects contain vanadium. For unusual bleggs, it may be a different percentage: 95% of purple bleggs contain vanadium, 92% of hard bleggs contain vanadium, etc.

Now suppose you find a blue egg-shaped furred flexible opaque object, an ordinary blegg in every visible way, and just for kicks you take it to the sorting scanner, and the scanner says "palladium"—this is one of the rare 2%. Is it a blegg?

At first you might answer that, since you intend to throw this object in the rube bin, you might as well call it a "rube". However, it turns out that almost all bleggs, if you switch off the lights, glow faintly in the dark; while almost all rubes do not glow in the dark. And the percentage of bleggs that glow in the dark is not significantly different for blue egg-shaped furred flexible opaque objects that contain palladium, instead of vanadium. Thus, if you want to guess whether the object glows like a blegg, or remains dark like a rube, you should guess that it glows like a blegg.

So is the object *really* a blegg or a rube?

On one hand, you'll throw the object in the rube bin no matter what else you learn. On the other hand, if there are any unknown characteristics of the object you need to infer, you'll infer them as if the object were a blegg, not a rube—group it into the similarity cluster of blue egg-shaped furred flexible opaque things, and not the similarity cluster of red cube-shaped smooth hard translucent things.

The question “Is this object a blegg?” may stand in for different queries on different occasions.

If it weren't standing in for *some* query, you'd have no reason to care.

Is atheism a “religion”<sup>3</sup>? Is transhumanism a “cult”<sup>4</sup>? People who argue that atheism is a religion “because it states beliefs about God” are really trying to argue (I think) that the reasoning methods used in atheism are on a par with the reasoning methods used in religion, or that atheism is no safer than religion in terms of the probability of causally engendering violence, etc... What's really at stake is an atheist's claim of substantial difference and superiority relative to religion, which the religious person is trying to reject by denying the difference rather than the superiority(!)

But that's not the a priori irrational part: The a priori irrational part is where, in the course of the argument, someone pulls out a dictionary and looks up the definition of “atheism” or “religion”. (And yes, it's just as silly whether an atheist or religionist does it.) How could a dictionary *possibly* decide whether an empirical cluster of atheists is really substantially different from an empirical cluster of theologians? How can reality vary with the meaning of a word? The points in thingspace don't move around when we re-draw a boundary.

But people often don't *realize* that their argument about where to draw a definitional boundary, is really a dispute over whether to infer a characteristic shared by most things inside an empirical cluster...

Hence the phrase, “disguised query”.

---

3. Page 392, ‘Uncritical Supercriticality’.

4. Page 432, ‘Cultish Countercultishness’.

## 9. Neural Categories<sup>1</sup>

**Followup to:** Disguised Queries<sup>2</sup>

In Disguised Queries<sup>3</sup>, I talked about a classification task of “bleggs” and “rubes”. The typical blegg is blue, egg-shaped, furred, flexible, opaque, glows in the dark, and contains vanadium. The typical rube is red, cube-shaped, smooth, hard, translucent, unglowing, and contains palladium. For the sake of simplicity, let us forget the characteristics of flexibility/hardness and opaqueness/translucency. This leaves five dimensions in thingspace<sup>4</sup>: Color, shape, texture, luminance, and interior.

Suppose I want to create an Artificial Neural Network (ANN) to predict unobserved blegg characteristics from observed blegg characteristics. And suppose I’m fairly naive about ANNs: I’ve read excited popular science books about how neural networks are distributed, emergent, and parallel *just like the human brain!!* but I can’t derive the differential equations for gradient descent in a non-recurrent multilayer network with sigmoid units (which is actually a lot easier than it sounds).

Then I might design a neural network that looks something like this:

---

1. [http://lesswrong.com/lw/nm/neural\\_categories/](http://lesswrong.com/lw/nm/neural_categories/)

2. Page 213, ‘Disguised Queries’.

3. Page 213, ‘Disguised Queries’.

4. Page 209, ‘The Cluster Structure of Thingspace’.

### Bleggi\_3

Network 1 is for classifying bleggs and rubes. But since “blegg” is an unfamiliar and synthetic concept, I’ve also included a similar Network 1b for distinguishing humans from Space Monsters, with input from Aristotle (“All men are mortal”) and Plato’s Academy (“A featherless biped with broad nails”).

A neural network needs a learning rule. The obvious idea is that when two nodes are often active at the same time, we should strengthen the connection between them—this is one of the first rules ever proposed for training a neural network, known as Hebb’s Rule.

Thus, if you often saw things that were both blue and furred—thus simultaneously activating the “color” node in the + state and the “texture” node in the + state—the connection would strengthen between color and texture, so that + colors activated + textures, and vice versa. If you saw things that were blue and egg-shaped and vanadium-containing, that would strengthen positive mutual connections between color and shape and interior.

Let’s say you’ve already seen plenty of bleggs and rubes come off the conveyor belt. But now you see something that’s furred, egg-shaped, and—gasp!—reddish purple (which we’ll model as a “color”

activation level of  $-2/3$ ). You haven't yet tested the luminance, or the interior. What to predict, what to predict?

What happens then is that the activation levels in Network 1 bounce around a bit. Positive activation flows luminance from shape, negative activation flows to interior from color, negative activation flows from interior to luminance... Of course all these messages are passed in *parallel!!* and *asynchronously!!* just like the human brain...

Finally Network 1 settles into a stable state, which has high positive activation for "luminance" and "interior". The network may be said to "expect" (though it has not yet seen) that the object will glow in the dark, and that it contains vanadium.

And lo, Network 1 exhibits this behavior even though there's no explicit node that says whether the object is a blegg or not. The judgment is *implicit in the whole network!!* Bleggness is an *attractor!!* which arises as the result of *emergent behavior!!* from the *distributed!!* learning rule.

Now in real life, this kind of network design—however faddish<sup>5</sup> it may sound—runs into *all sorts* of problems. Recurrent networks don't always settle right away: They can oscillate, or exhibit chaotic behavior, or just take a very long time to settle down. This is a Bad Thing when you see something big and yellow and striped, and you have to wait five minutes for your distributed neural network to settle into the "tiger" attractor. Asynchronous and parallel it may be, but it's not real-time.

And there are other problems, like double-counting the evidence<sup>6</sup> when messages bounce back and forth: If you suspect that an object glows in the dark, your suspicion will activate belief that the object contains vanadium, which in turn will activate belief that the object glows in the dark.

Plus if you try to scale up the Network 1 design, it requires  $O(N^2)$  connections, where N is the total number of observables.

So what might be a more realistic neural network design?

---

5. [http://thedailywtf.com/Articles/No,\\_We\\_Need\\_a\\_Neural\\_Network.aspx](http://thedailywtf.com/Articles/No,_We_Need_a_Neural_Network.aspx)

6. Page 147, 'Fake Causality'.

## Blegg2

In this network, a wave of activation converges on the central node from any clamped (observed) nodes, and then surges back out again to any unclamped (unobserved) nodes. Which means we can compute the answer in one step, rather than waiting for the network to settle—an important requirement in biology when the neurons only run at 20Hz. And the network architecture scales as  $O(N)$ , rather than  $O(N^2)$ .

Admittedly, there are some things you can notice more easily with the first network architecture than the second. Network 1 has a direct connection between every two nodes. So if red objects *never* glow in the dark, but red furred objects usually have the other blegg characteristics like egg-shape and vanadium, Network 1 can easily represent this: it just takes a very strong direct negative connection from color to luminance, but more powerful positive connections from texture to all other nodes except luminance.

Nor is this a “special exception” to the general rule that bleggs glow—remember, in Network 1, there is no unit that represents blegg-ness; blegg-ness emerges as an attractor in the distributed network.

So yes, those  $N^2$  connections were buying us something. But not very much. Network 1 is not *more* useful on most real-world

problems, where you rarely find an animal stuck halfway between being a cat and a dog.

(There are also facts that you can't easily represent in Network 1 or Network 2. Let's say sea-blue color and spheroid shape, when found together, always indicate the presence of palladium; but when found individually, without the other, they are each very strong evidence for vanadium. This is hard to represent, in either architecture, without extra nodes. Both Network 1 and Network 2 embody implicit assumptions about what kind of environmental structure is likely to exist; the ability to read this off is what separates the adults from the babes, in machine learning.)

Make no mistake: Neither Network 1, nor Network 2, are biologically realistic. *But* it still seems like a fair guess that however the brain really works, it is in some sense closer to Network 2 than Network 1. Fast, cheap, scalable, works well to distinguish dogs and cats: natural selection goes for that sort of thing like water running down a fitness landscape.

It seems like an ordinary enough task to classify objects as either bleggs or rubes, tossing them into the appropriate bin. But would you notice if sea-blue objects never glowed in the dark?

Maybe, if someone presented you with twenty objects that were alike only in being sea-blue, and then switched off the light, and none of the objects glowed. If you got hit over the head with it, in other words. Perhaps by presenting you with all these sea-blue objects in a group, your brain forms a new subcategory, and can detect the “doesn’t glow” characteristic within that subcategory. But you probably wouldn’t notice if the sea-blue objects were scattered among a hundred other bleggs and rubes. It wouldn’t be *easy* or *intuitive* to notice, the way that distinguishing cats and dogs is easy and intuitive.

Or: “Socrates is human, all humans are mortal, therefore Socrates is mortal.” How did Aristotle know that Socrates was human? Well, Socrates had no feathers, and broad nails, and walked upright, and spoke Greek, and, well, was generally shaped like a human and acted like one. So the brain decides, once and for all, that Socrates is human; and from there, infers that Socrates is mortal like all other humans thus yet observed. It doesn’t seem easy or intuitive to ask how much wearing clothes, as opposed to using lan-

guage, is associated with mortality. Just, “things that wear clothes and use language are human” and “humans are mortal”.

Are there biases associated with trying to classify things into categories once and for all? Of course there are. See e.g. Cultish Countercultishness<sup>7</sup>.

To be continued...

---

7. Page 432, ‘Cultish Countercultishness’.

## 10. How An Algorithm Feels From Inside<sup>1</sup>

**Followup to:** Neural Categories<sup>2</sup>

“If a tree falls in the forest, and no one hears it, does it make a sound?” I remember seeing an actual argument get started on this subject—a fully naive argument that went nowhere near Berkeleyan subjectivism. Just:

“It makes a sound, just like any other falling tree!”

“But how can there be a sound that no one hears?”

The standard rationalist view would be that the first person is speaking as if “sound” means acoustic vibrations in the air; the second person is speaking as if “sound” means an auditory experience in a brain. If you ask “Are there acoustic vibrations?” or “Are there auditory experiences?”, the answer is at once obvious. And so the argument is really about the definition of the word “sound”.

I think the standard analysis is essentially correct. So let’s accept that as a premise, and ask: Why do people get into such an argument? What’s the underlying psychology?

A key idea of the heuristics and biases program is that mistakes are often more revealing of cognition than correct answers. Getting into a heated dispute about whether, if a tree falls in a deserted forest, it makes a sound, is traditionally considered a mistake.

So what kind of mind design corresponds to that error?

In Disguised Queries<sup>3</sup> I introduced the blegg/rube classification task, in which Susan the Senior Sorter explains that your job is to sort objects coming off a conveyor belt, putting the blue eggs or “bleggs” into one bin, and the red cubes or “rubes” into the rube bin. This, it turns out, is because bleggs contain small nuggets of vanadium ore, and rubes contain small shreds of palladium, both of which are useful industrially.

Except that around 2% of blue egg-shaped objects contain palladium instead. So if you find a blue egg-shaped thing that contains

---

1. [http://lesswrong.com/lw/no/how\\_an\\_algorithm\\_feels\\_from\\_inside/](http://lesswrong.com/lw/no/how_an_algorithm_feels_from_inside/)

2. Page 217, ‘Neural Categories’.

3. Page 213, ‘Disguised Queries’.

palladium, should you call it a “rube” instead? You’re going to put it in the rube bin—why not call it a “rube”?

But when you switch off the light, nearly all bleggs glow faintly in the dark. And blue egg-shaped objects that contain palladium are just as likely to glow in the dark as any other blue egg-shaped object.

So if you find a blue egg-shaped object that contains palladium, and you ask “Is it a blegg?”, the answer depends on what you have to do with the answer: If you ask “Which bin does the object go in?”, then you choose as if the object is a rube. But if you ask “If I turn off the light, will it glow?”, you predict as if the object is a blegg. In one case, the question “Is it a blegg?” stands in for the disguised query<sup>4</sup>, “Which bin does it go in?”. In the other case, the question “Is it a blegg?” stands in for the disguised query<sup>5</sup>, “Will it glow in the dark?”

Now suppose that you have an object that is blue and egg-shaped and contains palladium; and you have already observed that it is furred, flexible, opaque, and glows in the dark.

This answers *every* query, observes every observable introduced. There’s nothing left for a disguised query to stand *for*.

So why might someone feel an impulse to go on arguing whether the object is *really* a blegg?

---

4. Page 213, ‘Disguised Queries’.

5. Page 213, ‘Disguised Queries’.

## Blegg3

This diagram from Neural Categories<sup>6</sup> shows two different neural networks that might be used to answer questions about bleggs and rubes. Network 1 has a number of disadvantages—such as potentially oscillating/chaotic behavior, or requiring  $O(N^2)$  connections—but Network 1's structure does have one major advantage over Network 2: Every unit in the network corresponds to a testable query. If you observe every observable, clamping every value, there are no units in the network left over.

Network 2, however, is a far better candidate for being something vaguely like how the human brain works: It's fast, cheap, scalable—and has an extra dangling unit in the center, whose activation can still vary, even after we've observed every single one of the surrounding nodes.

Which is to say that even after you know whether an object is blue or red, egg or cube, furred or smooth, bright or dark, and whether it contains vanadium or palladium, it *feels* like there's a leftover, unanswered question: *But is it really a blegg?*

Usually, in our daily experience, acoustic vibrations and auditory experience go together. But a tree falling in a deserted forest unbundles this common association. And even after you know that

---

6. Page 217, 'Neural Categories'.

the falling tree creates acoustic vibrations but not auditory experience, it *feels* like there's a leftover question: *Did it make a sound?*

We know where Pluto is, and where it's going; we know Pluto's shape, and Pluto's mass—but is it a planet?

Now remember: When you look at Network 2, as I've laid it out here, you're seeing the algorithm from the outside. People don't think to themselves, "Should the central unit fire, or not?" any more than you think "Should neuron #12,234,320,242 in my visual cortex fire, or not?"

It takes a deliberate effort to visualize your brain from the outside—and then you still don't see your actual brain; you imagine what you *think* is there, hopefully based on science, but regardless, you don't have any direct access to neural network structures from introspection. That's why the ancient Greeks didn't invent computational neuroscience.

When you look at Network 2, you are seeing from the *outside*; but the way that neural network structure feels from the *inside*, if you yourself *are* a brain running that algorithm, is that even after you know every characteristic of the object, you still find yourself wondering: "But is it a blegg, or not?"

This is a great gap to cross, and I've seen it stop people in their tracks. Because we don't instinctively see our intuitions as "intuitions", we just see them as the world. When you look at a green cup, you don't think of yourself as seeing a picture reconstructed in your visual cortex—although that *is* what you are seeing—you just see a green cup. You think, "Why, look, this cup is green," not, "The picture in my visual cortex of this cup is green."

And in the same way, when people argue over whether the falling tree makes a sound, or whether Pluto is a planet, they don't see themselves as arguing over whether a categorization should be active in their neural networks. It seems like either the tree makes a sound, or not.

We know where Pluto is, and where it's going; we know Pluto's shape, and Pluto's mass—but is it a planet? And yes, there were people who said this was a fight over definitions—but even that is a Network 2 sort of perspective, because you're arguing about how the central unit ought to be wired up. If you were a mind construct-

ed along the lines of Network 1, you wouldn't say "It depends on how you define 'planet,'" you would just say, "Given that we know Pluto's orbit and shape and mass, there is no question left to ask." Or, rather, that's how it would *feel*—it would *feel* like there was no question left—if you were a mind constructed along the lines of Network 1.

Before you can question your intuitions, you have to realize that what your mind's eye is looking at *is* an intuition—some cognitive algorithm, as seen from the inside—rather than a direct perception of the Way Things Really Are.

People cling to their intuitions<sup>7</sup>, I think, not so much because they believe their cognitive algorithms are perfectly reliable, but because they can't see their intuitions *as the way their cognitive algorithms happen to look from the inside*.

And so everything you try to say about how the native cognitive algorithm goes astray, ends up being contrasted to their direct perception of the Way Things Really Are—and discarded as obviously wrong.

---

7. [http://lesswrong.com/lw/n1/allais\\_malaise/](http://lesswrong.com/lw/n1/allais_malaise/)

## 11. Disputing Definitions<sup>1</sup>

### Followup to: How An Algorithm Feels From Inside<sup>2</sup>

I have watched more than one conversation—even conversations supposedly about cognitive science—go the route of disputing over definitions. Taking the classic example to be “If a tree falls in a forest, and no one hears it, does it make a sound?”, the dispute often follows a course like this:

*If a tree falls in the forest, and no one hears it, does it make a sound?*

Albert: “Of course it does. What kind of silly question is that? Every time I’ve listened to a tree fall, it made a sound, so I’ll guess that other trees falling also make sounds. I don’t believe the world changes around when I’m not looking.”

Barry: “Wait a minute. If no one hears it, how can it be a sound?”

In this example, Barry is arguing with Albert because of a genuinely different intuition about what constitutes a sound. But there’s more than one way the Standard Dispute can start. Barry could have a motive for rejecting Albert’s conclusion. Or Barry could be a skeptic who, upon hearing Albert’s argument, reflexively scrutinized<sup>3</sup> it for possible logical flaws; and then, on finding a counterargument, automatically accepted<sup>4</sup> it without applying a second layer of search for a counter-counterargument; thereby arguing himself into the opposite position. This doesn’t require that Barry’s *prior* intuition—the intuition Barry would have had, if we’d asked him before Albert spoke—have differed from Albert’s.

Well, if Barry didn’t have a differing intuition before, he sure has one now.

---

1. [http://lesswrong.com/lw/np/disputing\\_definitions/](http://lesswrong.com/lw/np/disputing_definitions/)

2. Page 223, ‘How An Algorithm Feels From Inside’.

3. Page 505, ‘Knowing About Biases Can Hurt People’.

4. Page 529, ‘Motivated Stopping and Motivated Continuation’.

Albert: “What do you mean, there’s no sound? The tree’s roots snap, the trunk comes crashing down and hits the ground. This generates vibrations that travel through the ground and the air. That’s where the energy of the fall goes, into heat and sound. Are you saying that if people leave the forest, the tree violates conservation of energy?”

Barry: “But no one *hears* anything. If there are no humans in the forest, or, for the sake of argument, anything else with a complex nervous system capable of ‘hearing’, then no one hears a sound.”

Albert and Barry recruit arguments that *feel* like support for their respective positions, describing in more detail the thoughts that caused their “sound”-detectors<sup>5</sup> to fire or stay silent. But so far the conversation has still focused on the forest, rather than definitions. And note that they don’t actually disagree on anything that happens in the forest.

Albert: “This is the dumbest argument I’ve ever been in. You’re a niddlewicking fallumphing pickleplumber.”

Barry: “Yeah? Well, you look like your face caught on fire and someone put it out with a shovel.”

Insult has been proffered and accepted; now neither party can back down without losing face. Technically, this isn’t part of the *argument*, as rationalists account such things; but it’s such an important part of the Standard Dispute that I’m including it anyway.

Albert: “The tree produces acoustic vibrations. By definition, that is a sound.”

Barry: “No one hears anything. By definition, that is not a sound.”

The argument starts shifting to focus on definitions. Whenever you feel tempted to say the words “by definition” in an argument

<sup>5</sup>. Page 223, ‘How An Algorithm Feels From Inside’.

that is not literally about pure mathematics, remember that anything which is true “by definition” is true in all possible worlds<sup>6</sup>, and so observing its truth can never constrain<sup>7</sup> *which* world you live in.

Albert: “My computer’s microphone can record a sound without anyone being around to hear it, store it as a file, and it’s called a ‘sound file’. And what’s stored in the file is the pattern of vibrations in air, not the pattern of neural firings in anyone’s brain. ‘Sound’ means a pattern of vibrations.”

Albert deploys an argument that *feels* like support for the word “sound” *having a particular meaning*. This is a different kind of question from whether acoustic vibrations take place in a forest—but the shift usually passes unnoticed.

Barry: “Oh, yeah? Let’s just see if the dictionary agrees with you.”

There’s a lot of things I could be curious about in the falling-tree scenario. I could go into the forest and look at trees, or learn how to derive the wave equation for changes of air pressure, or examine the anatomy of an ear, or study the neuroanatomy of the auditory cortex. Instead of doing any of these things, I am to consult a dictionary, apparently. Why? Are the editors of the dictionary expert botanists, expert physicists, expert neuroscientists? Looking in an encyclopedia might make sense, but why a *dictionary*?

Albert: “Hah! Definition 2c in Merriam-Webster: ‘Sound: Mechanical radiant energy that is transmitted by longitudinal pressure waves in a material medium (as air).’”

Barry: “Hah! Definition 2b in Merriam-Webster: ‘Sound: The sensation perceived by the sense of hearing.’”

---

6. Page 193, ‘The Parable of Hemlock’.

7. Page 138, ‘Fake Explanations’.

Albert and Barry, chorus: “Consarned dictionary! This doesn’t help at all!”

Dictionary editors are historians of usage, not legislators of language. Dictionary editors find words in current usage, then write down the words next to (a small part of<sup>8</sup>) what people seem to mean by them. If there’s more than one usage, the editors write down more than one definition.

Albert: “Look, suppose that I left a microphone in the forest and recorded the pattern of the acoustic vibrations of the tree falling. If I played that back to someone, they’d call it a ‘sound’! That’s the common usage! Don’t go around making up your own wacky definitions!”

Barry: “One, I can define a word any way I like so long as I use it consistently. Two, the meaning I gave was *in* the dictionary. Three, who gave *you* the right to decide what is or isn’t common usage?”

There’s quite a lot of rationality errors in the Standard Dispute. Some of them I’ve already covered, and some of them I’ve yet to cover; likewise the remedies.

But for now, I would just like to point out—in a mournful sort of way—that Albert and Barry seem to agree on virtually every question of what is *actually* going on inside the forest, and yet it doesn’t seem to generate any feeling of agreement.

Arguing about definitions is a garden path; people wouldn’t go down the path if they saw at the outset where it led. If you asked Albert (Barry) why he’s still arguing, he’d probably say something like: “Barry (Albert) is trying to sneak in his own definition of ‘sound’, the scurvey scoundrel, to support his ridiculous point; and I’m here to defend the standard definition.”

But suppose I went back in time to before the start of the argument:

---

8. Page 200, ‘Extensions and Intensions’.

*(Eliezer appears from nowhere in a peculiar conveyance that looks just like the time machine from the original 'The Time Machine' movie.)*

Barry: "Gosh! A time traveler!"

Eliezer: "I am a traveler from the future! Hear my words! I have traveled far into the past—around fifteen minutes—"

Albert: "Fifteen *minutes*?"

Eliezer: "—to bring you this message!"

*(There is a pause of mixed confusion and expectancy.)*

Eliezer: "Do you think that 'sound' should be defined to require both acoustic vibrations (pressure waves in air) and also auditory experiences (someone to listen to the sound), or should 'sound' be defined as meaning only acoustic vibrations, or only auditory experience?"

Barry: "You went back in time to ask us *that*?"

Eliezer: "My purposes are my own! Answer!"

Albert: "Well... I don't see why it would matter. You can pick any definition so long as you use it consistently."

Barry: "Flip a coin. Er, flip a coin twice."

Eliezer: "Personally I'd say that if the issue arises, both sides should switch to describing the event in unambiguous lower-level constituents, like acoustic vibrations or auditory experiences. Or each side could designate a new word, like 'alberzle' and 'bargulum', to use for what they respectively used to call 'sound'; and then both sides could use the new words consistently. That way neither side has to back down or lose face, but they can still communicate. And of course you should

try to keep track, at all times, of some testable proposition that the argument is actually about. Does that sound right to you?"

Albert: "I guess..."

Barry: "Why are we talking about this?"

Eliezer: "To preserve your friendship against a contingency you will, now, never know. For the future has already changed!"

*(Eliezer and the machine vanish in a puff of smoke.)*

Barry: "Where were we again?"

Albert: "Oh, yeah: If a tree falls in the forest, and no one hears it, does it make a sound?"

Barry: "It makes an alberzle but not a bargulum. What's the next question?"

This remedy doesn't destroy *every* dispute over categorizations. But it destroys a substantial fraction.

## 12. Feel the Meaning<sup>1</sup>

### Followup to: Disputing Definitions<sup>2</sup>

When I hear someone say, “Oh, look, a butterfly,” the spoken phonemes “butterfly” enter my ear and vibrate on my ear drum, being transmitted to the cochlea, tickling auditory nerves that transmit activation spikes to the auditory cortex, where phoneme processing begins, along with recognition of words, and reconstruction of syntax (a by no means serial process), and all manner of other complications.

But at the end of the day, or rather, at the end of the second, I am primed to look where my friend is pointing and see a visual pattern that I will recognize as a butterfly; and I would be quite surprised to see a wolf instead.

My friend looks at a butterfly, his throat vibrates and lips move, the pressure waves travel invisibly through the air, my ear hears and my nerves transduce and my brain reconstructs, and lo and behold, I know what my friend is looking at. Isn’t that marvelous? If we didn’t know about the pressure waves in the air, it would be a tremendous discovery in all the newspapers: Humans are telepathic! Human brains can transfer thoughts to each other!

Well, we *are* telepathic, in fact; but magic isn’t exciting when it’s merely *real*, and all your friends can do it too<sup>3</sup>.

Think telepathy is simple? Try building a computer that will be telepathic with you. Telepathy, or “language”, or whatever you want to call our partial thought transfer ability, is more complicated than it looks.

But it would be quite inconvenient to go around thinking, “Now I shall partially transduce some features of my thoughts into a linear sequence of phonemes which will invoke similar thoughts in my conversational partner...”

So the brain hides the complexity—or rather, never represents it in the first place—which leads people to think some peculiar thoughts about words.

---

1. [http://lesswrong.com/lw/nq/feel\\_the\\_meaning/](http://lesswrong.com/lw/nq/feel_the_meaning/)

2. Page 228, ‘Disputing Definitions’.

3. Page 177, “Science” as Curiosity-Stopper’.

As I remarked earlier<sup>4</sup>, when a large yellow striped object leaps at me, I think “Yikes! A tiger!” not “Hm... objects with the properties of largeness, yellowness, and stripedness have previously often possessed the properties ‘hungry’ and ‘dangerous’, and therefore, although it is not logically necessary, *auughhhh CRUNCH CRUNCH GULP.*”

Similarly, when someone shouts “Yikes! A tiger!”, natural selection would not favor an organism that thought, “Hm... I have just heard the syllables ‘Tie’ and ‘Grr’ which my fellow tribe members associate with their internal analogues of my own *tiger* concept, and which they are more likely to utter if they see an object they categorize as *aiiiieee CRUNCH CRUNCH help it's got my arm CRUNCH GULP.*”

Con-            Blegg4-4  
 sidering  
 this as a  
 design  
 con-  
 straint on  
 the hu-  
 man  
 cognitive  
 architec-  
 ture<sup>5</sup>, you  
 wouldn’t  
 want *any*  
 extra  
 steps be-  
 tween  
 when  
 your au-  
 ditory  
 cortex  
 recog-  
 nizes the syllables “tiger”, and when the tiger concept gets activated.

---

4. Page 197, ‘Words as Hidden Inferences’.

5. Page 223, ‘How An Algorithm Feels From Inside’.

Going back to the parable of bleggs and rubes<sup>6</sup>, and the centralized network<sup>7</sup> that categorizes quickly and cheaply, you might visualize a direct connection running from the unit that recognizes the syllable “blegg”, to the unit at the center of the blegg network. The central unit, the blegg concept, gets activated almost as soon as you hear Susan the Senior Sorter say “Blegg!”

Or, for purposes of talking—which also shouldn’t take eons—as soon as you see a blue egg-shaped thing and the central blegg unit fires, you holler “Blegg!” to Susan.

And what that algorithm feels like from inside<sup>8</sup> is that the label, and the concept, are very nearly *identified*; the meaning *feels like* an intrinsic property of the word itself.

The cognoscenti will recognize this as yet another case of E. T. Jaynes’s “Mind Projection Fallacy”. It feels like a word *has a* meaning, as a property of the word itself; just like how redness is a property of a red apple, or mysteriousness is a property of a mysterious phenomenon<sup>9</sup>.

Indeed, on most occasions, the brain will not distinguish at all between the word and the meaning—only bothering to separate the two while learning a new language, perhaps. And even then, you’ll see Susan pointing to a blue egg-shaped thing and saying “Blegg！”, and you’ll think, *I wonder what “blegg” means*, and not, *I wonder what mental category Susan associates to the auditory label “blegg”*.

Consider, in this light, the part of the Standard Dispute of Definitions<sup>10</sup> where the two parties argue about what the word “sound” *really* means—the same way they might argue whether a particular apple is *really* red or green:

Albert: “My computer’s microphone can record a sound without anyone being around to hear it, store it as a file, and it’s called a ‘sound file’. And what’s stored in the file is the pattern of vibrations in air, not the pattern of neural firings in anyone’s brain. ‘Sound’ means a pattern of vibrations.”

---

6. Page 213, ‘Disguised Queries’.

7. Page 217, ‘Neural Categories’.

8. Page 223, ‘How An Algorithm Feels From Inside’.

9. Page 154, ‘Mysterious Answers to Mysterious Questions’.

10. Page 228, ‘Disputing Definitions’.

Barry: “Oh, yeah? Let’s just see if the dictionary agrees with you.”

Albert feels intuitively that the word “sound” *has a meaning* and that the meaning *is* acoustic vibrations. Just as Albert feels that a tree falling in the forest *makes a sound* (rather than causing an event that *matches the sound category*).

Barry likewise *feels* that:

```
sound.meaning == auditory experiences
forest.sound == false
```

Rather than:

```
myBrain.FindConcept("sound") ==
concept_AuditoryExperience
concept_AuditoryExperience.match(forest)
== false
```

Which is closer to what’s *really* going on; but humans have not evolved to know this, anymore than humans instinctively know the brain is made of neurons.

Albert and Barry’s conflicting intuitions provide the fuel for continuing the argument in the phase of arguing over what the word “sound” means—which *feels* like arguing over a fact like any other fact, like arguing over whether the sky is blue or green.

You may not even notice that anything has gone astray, until you try to perform the rationalist ritual of stating a testable experiment<sup>11</sup> whose result depends on the facts you’re so heatedly disputing...

---

<sup>11</sup>. Page 105, ‘Making Beliefs Pay Rent (in Anticipated Experiences)’.

## 13. The Argument from Common Usage<sup>1</sup>

**Followup to:** Feel the Meaning<sup>2</sup>

Part of the Standard Definitional Dispute<sup>3</sup> runs as follows:

Albert: “Look, suppose that I left a microphone in the forest and recorded the pattern of the acoustic vibrations of the tree falling. If I played that back to someone, they’d call it a ‘sound’! That’s the common usage! Don’t go around making up your own wacky definitions!”

Barry: “One, I can define a word any way I like so long as I use it consistently. Two, the meaning I gave was *in* the dictionary. Three, who gave *you* the right to decide what is or isn’t common usage?”

Not all definitional disputes progress as far as recognizing the notion of common usage. More often, I think, someone picks up a dictionary because they believe that words have meanings<sup>4</sup>, and the dictionary faithfully records what this meaning is. Some people even seem to believe that the dictionary *determines* the meaning—that the dictionary editors are the Legislators of Language. Maybe because back in elementary school, their authority-teacher<sup>5</sup> said that they had to obey the dictionary, that it was a mandatory rule rather than an optional one?

Dictionary editors read what other people write, and record what the words seem to mean; they are historians. The Oxford English Dictionary may be *comprehensive*, but never *authoritative*.

But surely there is a social imperative to use words in a commonly understood way? Does not our human telepathy, our valuable power of language, rely on mutual coordination to work? Perhaps we should voluntarily treat dictionary editors as supreme arbiters—even if *they* prefer to think of themselves as historians—in

---

1. [http://lesswrong.com/lw/nr/the\\_argument\\_from\\_common\\_usage/](http://lesswrong.com/lw/nr/the_argument_from_common_usage/)

2. Page 234, ‘Feel the Meaning’.

3. Page 228, ‘Disputing Definitions’.

4. Page 234, ‘Feel the Meaning’.

5. Page 604, ‘Absolute Authority’.

order to maintain the quiet cooperation on which all speech depends.

The phrase “authoritative dictionary” is almost never used correctly, an example of proper usage being the Authoritative Dictionary of IEEE Standards. The IEEE is a body of voting members who have a professional need for exact agreement on terms and definitions, and so the Authoritative Dictionary of IEEE Standards is actual, negotiated legislation, which exerts whatever authority one regards as residing in the IEEE.

In everyday life, shared language usually does not arise from a deliberate agreement, as of the IEEE. It’s more a matter of infection, as words are invented and diffuse through the culture. (A “meme”, one might say, following Richard Dawkins thirty years ago—but you already know what I mean, and if not, you can look it up on Google, and then you too will have been infected.)

Yet as the example of the IEEE shows, agreement on language can also be a cooperatively established public good. If you and I wish to undergo an exchange of thoughts via language, the human telepathy, then it is in our mutual interest that we use the *same* word for similar concepts—preferably, concepts similar to the limit of resolution in our brain’s representation thereof—even though we have no obvious mutual interest in using any *particular* word for a concept.

We have no obvious mutual interest in using the word “oto” to mean sound, or “sound” to mean oto; but we have a mutual interest in using the *same* word, whichever word it happens to be. (Preferably, words we use frequently should be short, but let’s not get into information theory just yet.)

But, while we have a mutual interest, it is not strictly *necessary* that you and I use the similar labels *internally*; it is only convenient. If I know that, to you, “oto” means sound—that is, you associate “oto” to a concept very similar to the one I associate to “sound”—then I can say “Paper crumpling makes a crackling oto.” It requires extra thought, but I can do it if I want.

Similarly, if you say “What is the walking-stick of a bowling ball dropping on the floor?” and I know which concept *you* associate with the syllables “walking-stick”, then I can figure out what you mean. It may require some thought, and give me pause, because I

ordinarily associate “walking-stick” with a different concept. But I can do it just fine.

When humans really *want* to communicate with each other, we’re hard to stop! If we’re stuck on a deserted island with no common language, we’ll take up sticks and draw pictures in sand.

Albert’s appeal to the Argument from Common Usage assumes that agreement on language is a cooperatively established public good. Yet Albert assumes this for the sole purpose of rhetorically accusing Barry of breaking the agreement, and endangering the public good. Now the falling-tree argument has gone all the way from botany to semantics to politics; and so Barry responds by challenging Albert for the authority to define the word.

A rationalist, with the discipline of hugging the query<sup>6</sup> active, would notice that the conversation had gone rather far astray.

Oh, dear reader, is it all really necessary? Albert knows what Barry means by “sound”. Barry knows what Albert means by “sound”. Both Albert and Barry have access to words, such as “acoustic vibrations” or “auditory experience”, which they already associate to the same concepts, and which can describe events in the forest without ambiguity. If they were stuck on a deserted island, trying to communicate with each other, their work would be done.

When both sides *know* what the other side *wants* to say, and both sides accuse the other side of defecting from “common usage”, then whatever it is they are about, it is clearly not *working out a way to communicate with each other*. But this is the whole benefit that common usage provides in the first place.

Why would you argue about the meaning of a word, two sides trying to wrest it back and forth? If it’s just a namespace conflict that has gotten blown out of proportion, and nothing more is at stake, then the two sides need merely generate two new words and use them consistently<sup>7</sup>.

Yet often categorizations function as hidden inferences<sup>8</sup> and disguised queries<sup>9</sup>. Is atheism a “religion”<sup>10</sup>? If someone is arguing

---

6. Page 343, ‘Hug the Query’.

7. Page 228, ‘Disputing Definitions’.

8. Page 197, ‘Words as Hidden Inferences’.

9. Page 213, ‘Disguised Queries’.

10. Page 392, ‘Uncritical Supercriticality’.

that the reasoning methods used in atheism are on a par with the reasoning methods used in Judaism, or that atheism is on a par with Islam in terms of causally engendering violence, then they have a clear argumentative stake in lumping it all together into an indistinct gray blur of “faith<sup>11</sup>”.

Or consider the fight to blend together blacks and whites as “people”. This would not be a time to generate two words—what’s at stake is exactly the idea that you shouldn’t draw a moral distinction.

But once any empirical proposition is at stake, *or* any moral proposition, you can no longer appeal to common usage.

If the question is how to cluster together similar things<sup>12</sup> for purposes of inference, empirical predictions will depend on the answer; which means that definitions can be *wrong*. A conflict of predictions cannot be settled by an opinion poll.

If you want to know whether atheism should be clustered with supernaturalist religions for purposes of some particular empirical inference, the dictionary can’t answer you.

If you want to know whether blacks are people, the dictionary can’t answer you.

If everyone believes that the red light in the sky is Mars the God of War, the dictionary will define “Mars” as the God of War<sup>13</sup>. If everyone believes that fire is the release of phlogiston, the dictionary will define “fire” as the release of phlogiston.

There is an art to using words; even when definitions are not literally true or false, they are often wiser or more foolish. Dictionaries are mere histories of past usage; if you treat them as supreme arbiters of meaning, it binds you to the wisdom of the past<sup>14</sup>, forbidding you to do better<sup>15</sup>.

Though do take care to ensure (if you must depart from the wisdom of the past) that people can figure out what you’re trying to swim.

<sup>11</sup>. Page 599, ‘The Fallacy of Gray’.

<sup>12</sup>. Page 204, ‘Similarity Clusters’.

<sup>13</sup>. Page 200, ‘Extensions and Intensions’.

<sup>14</sup>. Page 409, ‘Guardians of the Truth’.

<sup>15</sup>. [http://lesswrong.com/lw/b9/tsuyoku\\_vs\\_the\\_egalitarian\\_instinct/](http://lesswrong.com/lw/b9/tsuyoku_vs_the_egalitarian_instinct/)

## 14. Empty Labels<sup>1</sup>

**Followup to:** The Argument from Common Usage<sup>2</sup>

Consider (yet again) the Aristotelian idea of categories. Let's say that there's some object with properties A, B, C, D, and E, or at least it looks E-ish.

Fred: "You mean that thing over there is blue, round, fuzzy, and—"

Me: "In Aristotelian logic, it's not supposed to make a difference what the properties are, or what I call them. That's why I'm just using the letters."

Next, I invent the Aristotelian category "zawa", which describes those objects, all those objects, and only those objects, which have properties A, C, and D.

Me: "Object 1 is zawa, B, and E."

Fred: "And it's blue—I mean, A—too, right?"

Me: "That's implied when I say it's zawa."

Fred: "Still, I'd like you to say it explicitly."

Me: "Okay. Object 1 is A, B, zawa, and E."

Then I add another word, "yokie", which describes all and only objects that are B and E; and the word "xippo", which describes all and only objects which are E but not D.

Me: "Object 1 is zawa and yokie, but not xippo."

Fred: "Wait, is it luminescent? I mean, is it E?"

Me: "Yes. That is the only possibility on the information given."

Fred: "I'd rather you spelled it out."

Me: "Fine: Object 1 is A, zawa, B, yokie, C, D, E, and not xippo."

Fred: "Amazing! You can tell all that just by looking?"

---

1. [http://lesswrong.com/lw/ns/empty\\_labels/](http://lesswrong.com/lw/ns/empty_labels/)

2. Page 238, 'The Argument from Common Usage'.

Impressive, isn't it? Let's invent even more new words: "Bolo" is A, C, and yokie; "mun" is A, C, and xippo; and "merlacdonian" is bolo and mun.

Pointlessly confusing? I think so too. Let's replace the labels with the definitions:

"Zawa, B, and E" becomes [A, C, D], B, E

"Bolo and A" becomes [A, C, [B, E]], A

"Merlacdonian" becomes [A, C, [B, E]], [A, C, [E, -D]]

And the thing to remember about the Aristotelian idea of categories is that [A, C, D] is the *entire* information of "zawa". It's not just that I can vary the label, but that I can get along just fine without any label at all—the rules for Aristotelian classes work purely on structures like [A, C, D]. To call one of these structures "zawa", or attach any other label to it, is a human convenience (or inconvenience) which makes not the slightest difference to the Aristotelian rules.

Let's say that "human" is to be defined as a mortal featherless biped. Then the classic syllogism<sup>3</sup> would have the form:

All [mortal, -feathers, bipedal] are mortal.

Socrates is a [mortal, -feathers, bipedal].

Therefore, Socrates is mortal.

The feat of reasoning looks a lot less impressive now, doesn't it?

Here the *illusion of inference* comes from the labels, which conceal the premises, and pretend to novelty in the conclusion. Replacing labels with definitions reveals the illusion, making visible the tautology's empirical unhelpfulness<sup>4</sup>. You can never say that Socrates is a [mortal, -feathers, biped] until you have observed him to be mortal.

There's an idea, which you may have noticed I hate, that "you can define a word any way you like". This idea came from the Aristotelian notion of categories; since, if you follow the Aristotelian rules *exactly* and *without flaw*—which humans never do<sup>5</sup>; Aristo-

---

3. Page 193, 'The Parable of Hemlock'.

4. Page 193, 'The Parable of Hemlock'.

5. Page 197, 'Words as Hidden Inferences'.

tle knew perfectly well that Socrates was human, even though that wasn't justified under his rules—but, *if* some imaginary nonhuman entity were to follow the rules exactly, they would never arrive at a contradiction. They wouldn't arrive at much of anything: they couldn't say that Socrates is a [mortal, ~feathers, biped] until they observed him to be mortal.

But it's not so much that labels are *arbitrary* in the Aristotelian system, as that the Aristotelian system works fine without *any labels at all*—it cranks out exactly the same stream of tautologies, they just look a lot less impressive. The labels are only there to create the *illusion* of inference.

So if you're going to have an Aristotelian proverb at all, the proverb should be, not “I can define a word any way I like,” nor even, “Defining a word never has any consequences,” but rather, “Definitions don't need words.”

## 15. Taboo Your Words<sup>1</sup>

### Followup to: Empty Labels<sup>2</sup>

In the game Taboo (by Hasbro), the objective is for a player to have their partner guess a word written on a card, without using that word or five additional words listed on the card. For example, you might have to get your partner to say “baseball” without using the words “sport”, “bat”, “hit”, “pitch”, “base” or of course “baseball”.

The existence of this game surprised me, when I discovered it. Why wouldn’t you just say “An artificial group conflict in which you use a long wooden cylinder to whack a thrown spheroid, and then run between four safe positions”?

But then, by the time I discovered the game, I’d already been practicing it for years—albeit with a different purpose.

Yesterday we saw how replacing terms with definitions could reveal the empirical unproductivity<sup>3</sup> of the classical Aristotelian syllogism:

All [mortal, -feathers, biped] are mortal;  
Socrates is a [mortal, -feathers, biped];  
Therefore Socrates is mortal.

But the principle applies much more broadly:

Albert: “A tree falling in a deserted forest makes a sound.”

Barry: “A tree falling in a deserted forest does not make a sound.”

Clearly, since one says “sound” and one says “-sound”, we must have a contradiction, right? But suppose that they both dereference their pointers before speaking:

Albert: “A tree falling in a deserted forest matches [membership test: this event generates acoustic

---

1. [http://lesswrong.com/lw/nu/taboo\\_your\\_words/](http://lesswrong.com/lw/nu/taboo_your_words/)

2. Page 242, ‘Empty Labels’.

3. Page 193, ‘The Parable of Hemlock’.

vibrations].”

Barry: “A tree falling in a deserted forest does not match [membership test: this event generates auditory experiences].”

Now there is no longer an apparent collision—all they had to do was prohibit themselves from using the word *sound*. If “acoustic vibrations” came into dispute, we would just play Taboo again and say “pressure waves in a material medium”; if necessary we would play Taboo again on the word “wave<sup>4</sup>” and replace it with the wave equation. (Play Taboo on “auditory experience” and you get “That form of sensory processing, within the human brain, which takes as input a linear time series of frequency mixes.”)

But suppose, on the other hand, that Albert and Barry were to have the argument:

Albert: “Socrates matches the concept [membership test: this person will die after drinking hemlock].”

Barry: “Socrates matches the concept [membership test: this person will not die after drinking hemlock].”

Now Albert and Barry have a substantive clash of expectations; a difference in what they anticipate seeing after Socrates drinks hemlock. But they might not notice this, if they happened to use the same word “human” for their different concepts.

You get a very different picture of what people agree or disagree about, depending on whether you take a label’s-eye-view (Albert says “sound” and Barry says “not sound”, so they must disagree) or taking the test’s-eye-view (Albert’s membership test is acoustic vibrations, Barry’s is auditory experience).

Get together a pack of *soi-disant* futurists and ask them if they believe we’ll have Artificial Intelligence in thirty years, and I would guess that at least half of them will say yes. If you leave it at that, they’ll shake hands and congratulate themselves on their consensus. But make the term “Artificial Intelligence” taboo, and ask them to describe *what* they expect to see, without ever using words like “computers” or “think”, and you might find quite a conflict of expectations hiding under that featureless standard word. Likewise

---

4. Page 140, ‘Guessing the Teacher’s Password’.

that other term<sup>5</sup>. And see also Shane Legg's compilation of 71 definitions of "intelligence"<sup>6</sup>.

The illusion of unity across religions can be dispelled by making the term "God" taboo, and asking them to say what it is they believe in; or making the word "faith" taboo, and asking them why they believe it. Though mostly they won't be able to answer at all, because it is mostly profession<sup>7</sup> in the first place, and you cannot cognitively zoom in on an audio recording.

When you find yourself in philosophical difficulties, the first line of defense is not to define your problematic terms, but to see whether you can think without using those terms at all. Or any of their short synonyms. And be careful not to let yourself invent a new word to use instead. Describe outward observables and interior mechanisms; don't use a single handle, whatever that handle may be.

Albert says that people have "free will". Barry says that people don't have "free will". Well, that will certainly generate an apparent conflict. Most philosophers would advise Albert and Barry to try to define exactly what they mean by "free will", on which topic they will certainly be able to discourse at great length. I would advise Albert and Barry to describe what it is that they think people do, or do not have, without using the phrase "free will" at all. (If you want to try this at home, you should also avoid the words "choose", "act", "decide", "determined", "responsible", or any of their synonyms.)

This is one of the nonstandard tools in my toolbox, and in my humble opinion, it works *way way* better than the standard one. It also requires more effort to use; you get what you pay for.

---

5. <http://www.singinst.org/blog/2007/09/30/three-major-singularity-schools/>

6. <http://arxiv.org/abs/0706.3639>

7. Page 108, 'Belief in Belief'.

## 16. Replace the Symbol with the Substance<sup>1</sup>

**Continuation of:** Taboo Your Words<sup>2</sup>

**Followup to:** Original Seeing<sup>3</sup>, Lost Purposes

4

What does it take to—as in yesterday’s example—see a “baseball game” as “An artificial group conflict in which you use a long wooden cylinder to whack a thrown spheroid, and then run between four safe positions”? What does it take to play the rationalist version of Taboo, in which the goal is not to find a synonym that isn’t on the card, but to find a way of describing without the standard concept-handle?

You have to visualize. You have to make your mind’s eye see the details, as though looking for the first time. You have to perform an Original Seeing<sup>5</sup>.

Is that a “bat”? No, it’s a long, round, tapering, wooden rod, narrowing at one end so that a human can grasp and swing it.

Is that a “ball”? No, it’s a leather-covered spheroid with a symmetrical stitching pattern, hard but not metal-hard, which someone can grasp and throw, or strike with the wooden rod, or catch.

Are those “bases”? No, they’re fixed positions on a game field, that players try to run to as quickly as possible because of their safety within the game’s artificial rules.

The chief obstacle to performing an original seeing is that your mind already has a nice neat summary, a nice little easy-to-use concept handle. Like the word “baseball”, or “bat”, or “base”. It takes an effort to stop your mind from sliding down the familiar path, the easy path, the path of least resistance, where the small featureless word rushes in and obliterates the details you’re trying to see. A word itself can have the destructive force of cliche<sup>6</sup>; a word itself can carry the poison of a cached thought<sup>7</sup>.

---

1. [http://lesswrong.com/lw/nv/replace\\_the\\_symbol\\_with\\_the\\_substance/](http://lesswrong.com/lw/nv/replace_the_symbol_with_the_substance/)

2. Page 245, ‘Taboo Your Words’.

3. Page 457, ‘Original Seeing’.

4. [http://lesswrong.com/lw/le/lost\\_purposes/](http://lesswrong.com/lw/le/lost_purposes/)

5. Page 457, ‘Original Seeing’.

6. Page 345, ‘Rationality and the English Language’.

7. Page 451, ‘Cached Thoughts’.

Playing the game of Taboo<sup>8</sup>—being able to describe without using the standard pointer/label/handle—is one of the *fundamental* rationalist capacities. It occupies the same primordial level as the habit of constantly asking “Why?” or “What does this belief make me anticipate?”

The art is closely related to:

- Pragmatism, because seeing in this way often gives you a much closer connection to anticipated experience<sup>9</sup>, rather than propositional belief<sup>10</sup>;
- Reductionism, because seeing in this way often forces you to drop down to a lower level of organization, look at the parts instead of your eye skipping over the whole;
- Hugging the query<sup>11</sup>, because words often distract you from the question you really want to ask;
- Avoiding cached thoughts<sup>12</sup>, which will rush in using standard words, so you can block them by tabooing standard words;
- The writer’s rule of “Show, don’t tell!”, which has power among rationalists;
- And not losing sight of your original purpose<sup>13</sup>.

How could tabooing a word help you keep your purpose?

From Lost Purposes<sup>14</sup>:

As you read this, some young man or woman is sitting at a desk in a university, earnestly studying material they have no intention of ever using, and no interest in knowing for its own sake. They want a high-paying job, and the high-paying job requires a piece of paper, and the piece of paper requires a previous master’s degree, and the master’s degree requires a bachelor’s degree, and the university that grants the bachelor’s degree requires you to take a class in 12th-century knitting patterns to graduate. So they diligently study, intending to forget it

---

8. Page 245, ‘Taboo Your Words’.

9. Page 105, ‘Making Beliefs Pay Rent (in Anticipated Experiences)’.

10. Page 108, ‘Belief in Belief’.

11. Page 343, ‘Hug the Query’.

12. Page 451, ‘Cached Thoughts’.

13. [http://lesswrong.com/lw/le/lost\\_purposes/](http://lesswrong.com/lw/le/lost_purposes/)

14. [http://lesswrong.com/lw/le/lost\\_purposes/](http://lesswrong.com/lw/le/lost_purposes/)

all the moment the final exam is administered, but still seriously working away, because they *want* that piece of paper.

Why are you going to “school”? To get an “education” ending in a “degree”. Blank out the forbidden words and all their obvious synonyms, visualize the actual details, and you’re much more likely to notice that “school” currently seems to consist of sitting next to bored teenagers listening to material you already know, that a “degree” is a piece of paper with some writing on it, and that “education” is forgetting the material as soon as you’re tested on it.

Leaky generalizations<sup>15</sup> often manifest through categorizations: People who actually learn in classrooms are categorized as “getting an education”, so “getting an education” must be good; but then anyone who actually shows up at a college will also match against the concept “getting an education”, whether or not they learn.

Students who understand math will do well on tests, but if you require schools to produce good test scores, they’ll spend all their time teaching to the test. A *mental category*, that imperfectly matches your goal, can produce the same kind of incentive failure *internally*. You want to learn, so you need an “education”; and then as long as you’re getting anything that matches against the category “education”, you may not notice whether you’re learning or not. Or you’ll notice, but you won’t realize you’ve lost sight of your original purpose, because you’re “getting an education” and that’s how you mentally described your goal.

To categorize is to throw away information. If you’re told that a falling tree makes a “sound”, you don’t know what the actual sound is; you haven’t actually heard the tree falling. If a coin lands “heads”, you don’t know its radial orientation. A blue egg-shaped thing may be a “blegg”, but what if the exact egg shape varies, or the exact shade of blue? You want to use categories to throw away irrelevant information, to sift gold from dust, but often the standard categorization ends up throwing out relevant information too. And when you end up in that sort of mental trouble, the first and most obvious solution is to play Taboo.

For example: “Play Taboo” is itself a leaky generalization. Hasbro’s version is not the rationalist version; they only list five ad-

<sup>15</sup> [http://lesswrong.com/lw/lo/leaky\\_generalizations/](http://lesswrong.com/lw/lo/leaky_generalizations/)

ditional banned words on the card, and that's not nearly enough coverage to exclude thinking in familiar old words. What rationalists do would count as playing Taboo—it would match against the “play Taboo” concept—but not everything that counts as playing Taboo works to force original seeing. If you just think “play Taboo to force original seeing”, you'll start thinking that anything that counts as playing Taboo must count as original seeing.

The rationalist version isn't a game, which means that you can't win by trying to be clever and stretching the rules. You have to play Taboo with a voluntary handicap: Stop yourself from using synonyms that aren't on the card. You also have to stop yourself from inventing a new simple word or phrase that functions as an equivalent mental handle to the old one. You are trying to zoom in on your map, not rename the cities; dereference the pointer, not allocate a new pointer; see the events as they happen, not rewrite the cliche in a different wording.

By visualizing the problem in more detail, you can see the lost purpose: Exactly what do you do when you “play Taboo”? What purpose does each and every part serve?

If you see your activities and situation originally, you will be able to originally see your goals as well. If you can look with fresh eyes, as though for the first time, you will see yourself doing things that you would never dream of doing if they were not habits.

Purpose is lost whenever the substance (learning, knowledge, health) is displaced by the symbol (a degree, a test score, medical care). To heal a lost purpose, or a lossy categorization, you must do the reverse:

Replace the symbol with the substance; replace the signifier with the signified; replace the property with the membership test; replace the word with the meaning; replace the label with the concept; replace the summary with the details; replace the proxy question with the real question; dereference the pointer; drop into a lower level of organization; mentally simulate the process instead of naming it; zoom in on your map.

“The Simple Truth<sup>16</sup>” was generated by an exercise of this discipline to describe “truth” on a lower level of organization, without invoking terms like “accurate”, “correct”, “represent”, “reflect”, “se-

---

16. <http://yudkowsky.net/bayes/truthb.html>

mantic”, “believe”, “knowledge”, “map”, or “real”. (And remember that the goal is not *really* to play Taboo—the word “true” appears in the text, but *not* to define truth. It would get a buzzer in Hasbro’s game, but we’re not *actually* playing that game. Ask yourself whether the document fulfilled its purpose, not whether it followed the rules.)

Bayes’s Rule itself describes “evidence” in pure math, without using words like “implies”, “means”, “supports”, “proves”, or “justifies”. Set out to *define* such philosophical terms, and you’ll just go in circles.

And then there’s the most important word of all to Taboo. I’ve often<sup>17</sup> warned<sup>18</sup> that you should be careful not to overuse it, or even avoid the concept<sup>19</sup> in certain cases. Now you know the real reason why. It’s not a bad subject to think about. But your true understanding is measured by your ability to describe what you’re doing and why, *without* using that word or any of its synonyms.

---

17. <http://yudkowsky.net/virtues/>

18. Page 422, ‘Two Cult Koans’.

19. [http://lesswrong.com/lw/nc/newcombs\\_problem\\_and\\_regret\\_of\\_rationality/](http://lesswrong.com/lw/nc/newcombs_problem_and_regret_of_rationality/)

## 17. Fallacies of Compression<sup>1</sup>

**Followup to:** Replace the Symbol with the Substance<sup>2</sup>

“The map is not the territory,” as the saying goes. The only life-size, atomically detailed, 100% accurate map of California is California. But California has important regularities, such as the shape of its highways, that can be described using vastly less information—not to mention vastly less *physical material*—than it would take to describe every atom within the state borders. Hence the *other* saying: “The map is not the territory, but you can’t fold up the territory and put it in your glove compartment.”

A paper map of California, at a scale of 10 kilometers to 1 centimeter (a million to one), doesn’t have room to show the distinct position of two fallen leaves lying a centimeter apart on the sidewalk. Even if the map tried to show the leaves, the leaves would appear as the same point on the map; or rather the map would need a feature size of 10 nanometers, which is a finer resolution than most book printers handle, not to mention human eyes.

Reality is very large—just the part we can see is billions of lightyears across. But your map of reality is written on a few pounds of neurons, folded up to fit inside your skull. I don’t mean to be insulting, but your skull is tiny, comparatively speaking.

Inevitably, then, certain things that are distinct in reality, will be compressed into the same point on your map.

But what this feels like from inside<sup>3</sup> is not that you say, “Oh, look, I’m compressing two things into one point on my map.” What it *feels* like from inside is that there is just *one* thing, and you are seeing it.

A sufficiently young child, or a sufficiently ancient Greek philosopher, would not know that there were such things as “acoustic vibrations” or “auditory experiences”. There would just be a single thing that happened when a tree fell; a single event called “sound”.

---

1. [http://lesswrong.com/lw/nw/fallacies\\_of\\_compression/](http://lesswrong.com/lw/nw/fallacies_of_compression/)

2. Page 248, ‘Replace the Symbol with the Substance’.

3. Page 223, ‘How An Algorithm Feels From Inside’.

To realize that there are *two* distinct events, underlying *one* point on your map, is an essentially *scientific* challenge—a big, difficult scientific challenge.

Sometimes fallacies of compression result from confusing two known things under the same label—you know about acoustic vibrations, and you know about auditory processing in brains, but you call them both “sound” and so confuse yourself. But the more dangerous fallacy of compression arises from having *no idea whatsoever* that two distinct entities even *exist*. There is just one mental folder in the filing system, labeled “sound”, and everything thought about “sound” drops into that one folder. It’s not that there are two folders with the same label; there’s just a single folder. By default, the map is compressed; why would the brain create two mental buckets where one would serve?

Or think of a mystery novel in which the detective’s critical insight is that one of the suspects has an identical twin. In the course of the detective’s ordinary work, his job is just to observe that Carol is wearing red, that she has black hair, that her sandals are leather—but all these are *facts about* Carol. It’s easy enough to question an individual fact, like `WearsRed(Carol)` or `BlackHair(Carol)`. Maybe `BlackHair(Carol)` is false. Maybe Carol dyes her hair. Maybe `BrownHair(Carol)`. But it takes a subtler detective to wonder if the Carol in `WearsRed(Carol)` and `BlackHair(Carol)`—the Carol file into which his observations drop—should be split into *two* files. Maybe there are two Carols, so that the Carol who wore red is not the same woman as the Carol who had black hair.

Here it is the very act of *creating* two different buckets that is the stroke of genius insight. ‘Tis easier to question one’s facts than one’s ontology.

The map of reality contained in a human brain, unlike a paper map of California, can expand dynamically when we write down more detailed descriptions. But what this feels like from inside is not so much zooming in on a map, as fissioning an indivisible atom—taking *one thing* (it felt like one thing) and splitting it into two or more things.

Often this manifests in the creation of new words, like “acoustic vibrations” and “auditory experiences” instead of just “sound”. Something about creating the new name seems to allocate the new

bucket. The detective is liable to start calling one of his suspects “Carol-2” or “the Other Carol” almost as soon as he realizes that there are two of them.

But expanding the map isn’t always as simple as generating new city names. It is a stroke of scientific insight to realize that such things as acoustic vibrations, or auditory experiences, even *exist*.

The obvious modern-day illustration would be words like “intelligence” or “consciousness”. Every now and then one sees a press release claiming that a research has “explained consciousness” because a team of neurologists investigated a 40Hz electrical rhythm that might have something to do with cross-modality binding of sensory information, or because they investigated the reticular activating system that keeps humans awake. That’s an extreme example, and the usual failures are more subtle, but they are of the same kind. The part of “consciousness” that people find most interesting is reflectivity, self-awareness, realizing that the person I see in the mirror is “me”; that and the hard problem of subjective experience as distinguished by Chalmers. We also label “conscious” the state of being awake, rather than asleep, in our daily cycle. But they are all different concepts going under the same name, and the underlying phenomena are different scientific puzzles. You can explain being awake without explaining reflectivity or subjectivity.

Fallacies of compression also underlie the bait-and-switch technique in philosophy—you argue about “consciousness” under one definition (like the ability to think about thinking) and then apply the conclusions to “consciousness” under a different definition (like subjectivity). Of course it may be that the two are the same thing, but if so, genuinely *understanding* this fact would require *first* a conceptual split and *then* a genius stroke of reunification.

Expanding your map is (I say again) a *scientific* challenge: part of the art of science, the skill of inquiring into the world. (And of course you cannot solve a scientific challenge by appealing to dictionaries, nor master a complex skill of inquiry by saying “I can define a word any way I like”.) Where you see a single confusing thing, with protean and self-contradictory attributes, it is a good guess that your map is cramming too much into one point—you need to pry it apart and allocate some new buckets. This is not like *defining* the single thing you see, but it *does* often follow from figuring out how to talk about the thing without using a single mental handle.

So the skill of prying apart the map is linked to the rationalist version<sup>4</sup> of Taboo<sup>5</sup>, and to the wise use of words; because words often represent the points on our map, the labels under which we file our propositions and the buckets into which we drop our information. Avoiding a single word, or allocating new ones, is often part of the skill of expanding the map.

---

4. Page 248, 'Replace the Symbol with the Substance'.  
5. Page 245, 'Taboo Your Words'.

## 18. Categorizing Has Consequences<sup>1</sup>

### Followup to: Fallacies of Compression<sup>2</sup>

Among the many genetic variations and mutations you carry in your genome, there are a very few alleles you probably know—including those determining your blood type: the presence or absence of the A, B, and + antigens. If you receive a blood transfusion containing an antigen you don't have, it will trigger an allergic reaction. It was Karl Landsteiner's discovery of this fact, and how to test for compatible blood types, that made it possible to transfuse blood without killing the patient. (1930 Nobel Prize in Medicine.) Also, if a mother with blood type A (for example) bears a child with blood type A+, the mother may acquire an allergic reaction to the + antigen; if she has another child with blood type A+, the child will be in danger, unless the mother takes an allergic suppressant during pregnancy. Thus people learn their blood types before they marry.

Oh, and *also*: people with blood type A are earnest and creative, while people with blood type B are wild and cheerful. People with type O are agreeable and sociable, while people with type AB are cool and controlled. (You would think that O would be the absence of A and B, while AB would just be A plus B, but no...) All this, according to the Japanese blood type theory of personality<sup>3</sup>. It would seem that blood type plays the role in Japan that astrological signs play in the West, right down to blood type horoscopes in the daily newspaper.

This fad is especially odd because blood types have *never been* mysterious, not in Japan and not anywhere. We only know blood types even *exist* thanks to Karl Landsteiner. No mystic witch doctor, no venerable sorcerer, ever said a word about blood types; there are no ancient, dusty scrolls to shroud the error in the aura of antiquity<sup>4</sup>. If the medical profession claimed tomorrow that it had all been a colossal hoax, we layfolk would not have one scrap of evidence from our unaided senses to contradict them.

---

1. [http://lesswrong.com/lw/nx/categorizing\\_has\\_consequences/](http://lesswrong.com/lw/nx/categorizing_has_consequences/)

2. Page 253, 'Fallacies of Compression'.

3. [http://en.wikipedia.org/wiki/Japanese\\_blood\\_type\\_theory\\_of\\_personality](http://en.wikipedia.org/wiki/Japanese_blood_type_theory_of_personality)

4. Page 409, 'Guardians of the Truth'.

There's never been a war between blood types. There's never even been a political conflict between blood types. The stereotypes must have arisen *strictly* from the *mere existence* of the labels.

Now, someone is bound to point out that this is a story of categorizing humans. Does the same thing happen if you categorize plants, or rocks, or office furniture? I can't recall reading about such an experiment, but of course, that doesn't mean one hasn't been done<sup>5</sup>. (I'd expect the chief difficulty of doing such an experiment would be finding a protocol that didn't mislead the subjects into thinking that, since the label was given you, it must be significant somehow.) So while I don't mean to update on imaginary evidence, I would predict a positive result for the experiment: I would expect them to find that mere labeling had power over all things, at least in the human imagination.

You can see this in terms of similarity clusters<sup>6</sup>: once you draw a boundary around a group, the mind starts trying to harvest similarities from the group. And unfortunately the human pattern-detectors seem to operate in such overdrive that we see patterns whether they're there or not; a weakly negative correlation can be mistaken for a strong positive one with a bit of selective memory.

You can see this in terms of neural algorithms<sup>7</sup>: creating a name for a set of things is like allocating a subnetwork to find patterns in them.

You can see this in terms of a compression fallacy<sup>8</sup>: things given the same name end up dumped into the same mental bucket, blurring them together into the same point on the map.

Or you can see this in terms of the boundless human ability to make stuff up out of thin air and believe it because no one can prove it's wrong<sup>9</sup>. As soon as you name the category, you can start making up stuff about it. The named thing doesn't have to be perceptible; it doesn't have to exist; it doesn't even have to be coherent.

And no, it's not just Japan: Here in the West, a blood-type-based diet book called Eat Right 4 Your Type<sup>10</sup> was a bestseller.

---

5. [http://lesswrong.com/lw/kj/no\\_one\\_knows\\_what\\_science\\_doesnt\\_know/](http://lesswrong.com/lw/kj/no_one_knows_what_science_doesnt_know/)

6. Page 204, 'Similarity Clusters'.

7. Page 217, 'Neural Categories'.

8. Page 253, 'Fallacies of Compression'.

9. Page 505, 'Knowing About Biases Can Hurt People'.

10. [http://en.wikipedia.org/wiki/Blood\\_type\\_diet](http://en.wikipedia.org/wiki/Blood_type_diet)

Any way you look at it, drawing a boundary in thingspace<sup>II</sup> is not a neutral act. Maybe a more cleanly designed, more purely Bayesian AI could ponder an arbitrary class and not be influenced by it. But you, a human, do not have that option. Categories are not static things in the context of a human brain; as soon as you actually think of them, they exert force on your mind. One more reason not to believe you can define a word any way you like.

---

II. Page 209, ‘The Cluster Structure of Thingspace’.

## 19. Sneaking in Connotations<sup>1</sup>

### Followup to: Categorizing Has Consequences<sup>2</sup>

Yesterday, we saw that in Japan, blood types have taken the place of astrology—if your blood type is AB, for example, you’re supposed to be “cool and controlled”.

So suppose we decided to invent a new word, “wiggin”, and *defined* this word to mean people with green eyes and black hair—

A green-eyed man with black hair walked into a restaurant.

“Ha,” said Danny, watching from a nearby table, “did you see that? A wiggin just walked into the room.

Bloody wiggins. Commit all sorts of crimes, they do.”

His sister Erda sighed. “You haven’t *seen* him commit any crimes, have you, Danny?”

“Don’t need to,” Danny said, producing a dictionary. “See, it says right here in the Oxford English Dictionary. ‘Wiggin. (i) A person with green eyes and black hair.’ He’s got green eyes and black hair, he’s a wiggin. You’re not going to argue with the Oxford English Dictionary, are you? *By definition*, a green-eyed black-haired person is a wiggin.”

“But you called him a wiggin,” said Erda. “That’s a nasty thing to say about someone you don’t even know. You’ve got no evidence that he puts too much ketchup on his burgers, or that as a kid he used his slingshot to launch baby squirrels.”

“But he *is* a wiggin,” Danny said patiently. “He’s got green eyes and black hair, right? Just you watch, as soon as his burger arrives, he’s reaching for the ketchup.”

The human mind passes from observed characteristics to inferred characteristics via the medium of words. In “All humans are mortal, Socrates is a human, therefore Socrates is mortal<sup>3</sup>”, the observed characteristics are Socrates’s clothes, speech, tool use, and

1. [http://lesswrong.com/lw/ny/sneaking\\_in\\_connotations/](http://lesswrong.com/lw/ny/sneaking_in_connotations/)

2. Page 257, ‘Categorizing Has Consequences’.

3. Page 193, ‘The Parable of Hemlock’.

generally human shape; the categorization is “human”; the inferred characteristic is poisonability by hemlock.

Of course there’s no hard distinction between “observed characteristics” and “inferred characteristics”. If you hear someone speak, they’re probably shaped like a human, all else being equal. If you see a human figure in the shadows, then *ceteris paribus* it can probably speak.

And yet some properties do tend to be more inferred than observed. You’re more likely to decide that someone is human, and will therefore burn if exposed to open flame, than carry through the inference the other way around.

If you look in a dictionary for the definition of “human”, you’re more likely to find characteristics like “intelligence” and “featherless biped”—characteristics that are useful for quickly eyeballing what is and isn’t a human—rather than the ten thousand connotations, from vulnerability to hemlock, to overconfidence, that we can infer from someone’s being human. Why? Perhaps dictionaries are intended to let you match up labels to similarity groups, and so are designed to quickly isolate clusters in thingspace. Or perhaps the big, distinguishing characteristics are the most salient, and therefore first to pop into a dictionary editor’s mind. (I’m not sure how aware dictionary editors are of what they *really* do.)

But the upshot is that when Danny pulls out his OED to look up “wiggin”, he sees listed only the first-glance characteristics that distinguish a wiggin: Green eyes and black hair. The OED doesn’t list the many minor *connotations* that have come to attach to this term, such as criminal proclivities, culinary peculiarities, and some unfortunate childhood activities.

How did those connotations get there in the first place? Maybe there was once a famous wiggin with those properties. Or maybe someone made stuff up at random, and wrote a series of bestselling books about it (*The Wiggin*, *Talking to Wiggins*, *Raising Your Little Wiggin*, *Wiggins in the Bedroom*). Maybe even the wiggins believe it now, and act accordingly. As soon as you call some people “wiggins”, the word will begin acquiring connotations.

But remember the Parable of Hemlock<sup>4</sup>: If we go by the logical class definitions, we can never class Socrates as a “human” until

---

4. Page 193, ‘The Parable of Hemlock’.

after we observe him to be mortal. Whenever someone pulls a dictionary, they're generally trying to sneak in a *connotation*, not the actual definition written down in the dictionary.

After all, if the *only* meaning of the word "wiggin" is "green-eyed black-haired person", then why not just call those people "green-eyed black-haired people"? And if you're wondering whether someone is a ketchup-reacher, why not ask<sup>5</sup> directly<sup>6</sup>, "Is he a ketchup-reacher?" rather than "Is he a wiggin?" (Note substitution of substance for symbol<sup>7</sup>.)

Oh, but arguing the *real* question would require *work*. You'd have to actually watch the wiggin to see if he reached for the ketchup. Or maybe see if you can find statistics on how many green-eyed black-haired people actually like ketchup. At any rate, you wouldn't be able to do it sitting in your living room with your eyes closed. And people are lazy. They'd rather argue "by definition", especially since they think "you can define a word any way you like".

But of course the *real* reason they care whether someone is a "wiggin" is a connotation—a feeling that comes along with the word—that isn't in the definition they *claim* to use.

Imagine Danny saying, "Look, he's got green eyes and black hair. He's a wiggin! It says so right there in the dictionary!—therefore, he's got black hair. Argue with that, if you can!"

Doesn't have much of a triumphant ring to it, does it? If the real point of the argument actually *was* contained in the dictionary definition—if the argument genuinely *was* logically valid—then the argument would *feel* empty; it would either say nothing new, or beg the question.

It's only the attempt to smuggle in connotations *not* explicitly listed in the definition, that makes anyone feel they can *score a point* that way.

---

5. Page 197, 'Words as Hidden Inferences'.

6. Page 343, 'Hug the Query'.

7. Page 248, 'Replace the Symbol with the Substance'.

## 20. Arguing “By Definition”<sup>1</sup>

### Followup to: Sneaking in Connotations<sup>2</sup>

“This plucked chicken has two legs and no feathers—therefore, *by definition*, it is a human!”

When people argue definitions, they usually start with some visible, known, or at least widely believed set of characteristics; then pull out a dictionary, and point out that these characteristics fit the dictionary definition; and so conclude, “Therefore, *by definition*, atheism is a religion!”

But visible, known, widely believed characteristics are rarely the real<sup>3</sup> point<sup>4</sup> of a dispute. Just the fact that someone thinks Socrates’s two legs are evident enough to make a good premise for the argument, “Therefore, *by definition*, Socrates is human!” indicates that bipedalism probably isn’t *really* what’s at stake—or the listener would reply, “Whaddaya mean Socrates is bipedal? That’s what we’re arguing about in the first place!”

Now there is an important sense in which we can legitimately move from evident characteristics to not-so-evident ones. You can, legitimately, see that Socrates is human-shaped, and predict his vulnerability to hemlock. But this *probabilistic* inference does not rely on dictionary definitions or common usage; it relies on the universe containing empirical clusters<sup>5</sup> of similar things<sup>6</sup>.

This cluster structure is not going to change depending on how you define your words. Even if you look up the dictionary definition of “human” and it says “all featherless bipeds except Socrates”, that isn’t going to change the *actual* degree to which Socrates is similar to the rest of us featherless bipeds.

When you are arguing *correctly* from cluster structure, you’ll say something like, “Socrates has two arms, two feet, a nose and tongue, speaks fluent Greek, uses tools, and in every aspect I’ve been able to observe him, seems to have every major and minor property that characterizes *Homo sapiens*; so I’m going to guess that he has human

1. [http://lesswrong.com/lw/nz/arguing\\_by\\_definition/](http://lesswrong.com/lw/nz/arguing_by_definition/)

2. Page 260, ‘Sneaking in Connotations’.

3. Page 213, ‘Disguised Queries’.

4. Page 260, ‘Sneaking in Connotations’.

5. Page 209, ‘The Cluster Structure of Thingspace’.

6. Page 204, ‘Similarity Clusters’.

DNA, human biochemistry, and is vulnerable to hemlock just like all other *Homo sapiens* in whom hemlock has been clinically tested for lethality.”

And suppose I reply, “But I saw Socrates out in the fields with some herbologists; I think they were trying to prepare an antidote. Therefore I *don't* expect Socrates to keel over after he drinks the hemlock—he will be an exception to the general behavior of objects in his cluster: they did not take an antidote, and he did.”

Now there's not much point in arguing over whether Socrates is “human” or not. The conversation has to move to a more detailed level<sup>7</sup>, poke around *inside* the details that make up the “human” category—talk about human biochemistry, and specifically, the neuro-toxic effects of coniine.

If you go on insisting, “But Socrates is a human and humans, *by definition*, are mortal!” then what you're really trying to do is blur out everything you know about Socrates *except* the fact of his humanity—insist that the only correct prediction is the one you would make if you knew nothing about Socrates *except* that he was human.

Which is like insisting that a coin is 50% likely to be showing heads or tails, because it is a “fair coin”, after you've *actually looked at the coin* and it's showing heads. It's like insisting that Frodo has ten fingers, because most hobbits have ten fingers, after you've *already looked at his hands* and seen nine fingers. Naturally this is illegal under Bayesian probability theory: You can't just refuse to condition on new evidence.

And you can't just keep one categorization and make estimates based on that, while deliberately throwing out everything else you know.

Not every piece of new evidence makes a significant difference, of course. If I see that Socrates has nine fingers, this isn't going to noticeably change my estimate of his vulnerability to hemlock, because I'll expect that the way Socrates lost his finger didn't change the rest of his biochemistry. And this is true, *whether or not* the dictionary's definition says that human beings have ten fingers. The legal inference is based on the cluster structure of the environment, and the causal structure of biology; *not* what the dictionary editor writes down, nor even “common usage”.

---

7. Page 248, ‘Replace the Symbol with the Substance’.

Now ordinarily, when you’re doing this *right*—in a *legitimate* way—you just say, “The coniine alkaloid found in hemlock produces muscular paralysis in humans, resulting in death by asphyxiation.” Or more simply, “Humans are vulnerable to hemlock.” That’s how it’s usually said in a *legitimate* argument.

When would someone feel the need to *strengthen* the argument with the emphatic phrase “by definition”? (I.e. “Humans are vulnerable to hemlock *by definition!*”) Why, when the inferred characteristic has been called into doubt—Socrates has been seen consulting herbologists—and so the speaker feels the need to tighten the vise of logic.

So when you see “by definition” used like this, it usually means: “Forget what you’ve heard about Socrates consulting herbologists—humans, *by definition*, are mortal!”

People feel the need to squeeze the argument onto a single course by saying “Any P, by definition, has property Q!”, on exactly those occasions when they see, and prefer to dismiss out of hand, *additional arguments* that call into doubt the default inference based on clustering.

So too with the argument “X, *by definition*, is a Y!” E.g., “Atheists believe that God doesn’t exist; therefore atheists have beliefs about God, because a negative belief is still a belief; therefore atheism asserts answers to theological questions; therefore atheism is, *by definition*, a religion.”

You wouldn’t feel the need to say, “Hinduism, *by definition*, is a religion!” because, well, of course Hinduism is a religion. It’s not just a religion “*by definition*”, it’s, like, an *actual* religion.

Atheism does not resemble the central members of the “religion” cluster, so if it wasn’t for the fact that atheism is a religion *by definition*, you might go around thinking that atheism *wasn’t* a religion. That’s why you’ve got to crush all opposition by pointing out that “Atheism is a religion” is true *by definition*, because it isn’t true any other way.

Which is to say: People insist that “X, *by definition*, is a Y!” on those occasions when they’re trying to sneak in a connotation<sup>8</sup> of Y that isn’t directly in the definition, and X doesn’t look all that much like other members of the Y cluster.

---

8. Page 260, ‘Sneaking in Connotations’.

Over the last thirteen years I've been keeping track of how often this phrase is used correctly versus incorrectly—though not with literal statistics, I fear. But eyeballing suggests that using the phrase *by definition*, anywhere outside of math, is among the most alarming signals of flawed argument I've ever found. It's right up there with "Hitler"<sup>9</sup>, "God", "absolutely certain" and "can't prove that".

This heuristic of failure is not perfect—the first time I ever spotted a correct usage outside of math, it was by Richard Feynman; and since then I've spotted more. But you're probably better off just deleting the phrase "by definition" from your vocabulary—and *always* on any occasion where you might be tempted to say it in italics or followed with an exclamation mark. That's a bad idea *by definition!*

---

9. <http://www.catb.org/jargon/html/G/Godwins-Law.html>

## 21. Where to Draw the Boundary?<sup>1</sup>

**Followup to:** Arguing “By Definition”<sup>2</sup>

The one comes to you and says<sup>3</sup>:

Long have I pondered the meaning of the word “Art”, and at last I’ve found what seems to me a satisfactory definition: “Art is that which is designed for the purpose of creating a reaction in an audience.”

*Just because there’s a word “art” doesn’t mean that it **has a meaning**, floating out there in the void, which you can **discover** by finding the right definition.*

It feels that way<sup>4</sup>, but it is not so.

Wondering how to *define a word* means you’re looking at the problem the wrong way—searching for the mysterious essence of what is, in fact, a communication signal<sup>5</sup>.

Now, there *is* a real challenge which a rationalist may legitimately attack, but the challenge is not to find a satisfactory definition of a word. The real challenge can be played as a single-player game, without speaking aloud. The challenge is figuring out which things are similar to each other—which things are clustered together—and sometimes, which things have a common cause.

If you define “electromugnetism” to include lightning, include compasses, exclude light, and include Mesmer’s “animal magnetism” (what we now call hypnosis), then you will have some trouble asking “How does electromugnetism work?” You have lumped together things which do not belong together, and excluded others that would be needed to complete a set. (This example is historically plausible; Mesmer came before Faraday.)

We could say that electromugnetism is a *wrong word*, a boundary in thingspace<sup>6</sup> that loops around and swerves through the clusters, a cut that fails to carve reality along its natural joints.

---

1. [http://lesswrong.com/lw/oo/where\\_to\\_draw\\_the\\_boundary/](http://lesswrong.com/lw/oo/where_to_draw_the_boundary/)

2. Page 263, ‘Arguing “By Definition”’.

3. [http://lesswrong.com/lw/nw/fallacies\\_of\\_compression/ia9](http://lesswrong.com/lw/nw/fallacies_of_compression/ia9)

4. Page 234, ‘Feel the Meaning’.

5. Page 238, ‘The Argument from Common Usage’.

6. Page 209, ‘The Cluster Structure of Thingspace’.

Figuring where to cut reality in order to carve along the joints—*this* is the problem worthy of a rationalist. It is what people *should* be trying to do, when they set out in search of the floating essence of a word.

And make no mistake: it is a *scientific* challenge to realize that you need a single word to describe breathing and fire<sup>7</sup>. So do not think to consult the dictionary editors, for that is not their job.

What is “art”? But there is no essence of the word, floating in the void.

Perhaps you come to me with a long list of the things that you call “art” and “not art”:

The *Little Fugue in G Minor*: Art.

A punch in the nose: Not art.

Escher’s *Relativity*: Art.

A flower: Not art.

The Python programming language: Art.

A cross floating in urine: Not art.

Jack Vance’s *Tschai* novels: Art.

Modern Art: Not art.

And you say to me: “It feels intuitive to me to draw this boundary, but I don’t know why—can you find me an intension that matches this extension? Can you give me a *simple* description of this boundary?”

So I reply: “I think it has to do with admiration of craftsmanship: work going in and wonder coming out. What the included items have in common is the similar aesthetic emotions that they inspire, and the deliberate human effort that went into them with the intent of producing such an emotion.”

Is this helpful, or is it just cheating<sup>8</sup> at Taboo<sup>9</sup>? I would argue that the list of which human emotions are or are not *aesthetic* is far more compact than the list of everything that is or isn’t art. You might be able to see those emotions lighting up an fMRI scan—I say this by way of emphasizing that emotions are not ethereal.

7. [http://lesswrong.com/lw/bq/universal\\_fire/](http://lesswrong.com/lw/bq/universal_fire/)

8. Page 248, ‘Replace the Symbol with the Substance’.

9. Page 245, ‘Taboo Your Words’.

But of course my definition of art is not the real point. The real point is that you could well dispute either the intension or the extension<sup>10</sup> of my definition.

You could say, “Aesthetic emotion is *not* what these things have in common; what they have in common is an intent to inspire *any* complex emotion for the sake of inspiring it.” That would be disputing my intension, my attempt to draw a curve through the data points. You would say, “Your equation may roughly fit those points, but it is not the true generating distribution.”

Or you could dispute my extension by saying, “Some of these things do belong together—I can see what you’re getting at—but the Python language shouldn’t be on the list, and Modern Art should be.” (This would mark you as a gullible philistine, but you could argue it.) Here, the presumption is that there is indeed an underlying curve that generates this apparent list of similar and dissimilar things—that there is a rhyme and reason, *even though you haven’t said yet where it comes from*—but I have unwittingly lost the rhythm and included some data points from a different generator.

Long before you *know* what it is that electricity and magnetism have in common, you might still suspect—based on surface appearances—that “animal magnetism” does not belong on the list.

Once upon a time it was thought that the word “fish” included dolphins. Now you could play the oh-so-clever arguer, and say, “The list: {Salmon, guppies, sharks, dolphins, trout} is just a list—you can’t say that a list is *wrong*. I can prove in set theory that this list exists. So my definition of *fish*, which is simply this extensional list, cannot possibly be ‘wrong’ as you claim.”

Or you could stop playing nitwit games and admit that dolphins don’t belong on the fish list.

You come up with a list of things that *feel* similar, and take a guess at why this is so. But when you finally discover what they *really* have in common, it may turn out that your guess was wrong. It may even turn out that your list was wrong.

You cannot hide behind a comforting shield of correct-by-definition. Both extensional definitions and intensional definitions can be wrong, can fail to carve reality at the joints.

---

<sup>10</sup> Page 200, ‘Extensions and Intensions’.

Categorizing is a guessing endeavor, in which you can make mistakes; so it's wise to be able to admit, from a theoretical standpoint, that your definition-guesses can be "mistaken".

## 22. Entropy, and Short Codes<sup>1</sup>

**Followup to:** Where to Draw the Boundary?<sup>2</sup>

Suppose you have a system X that's equally likely to be in any of 8 possible states:

$$\{X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8\}$$

There's an extraordinarily ubiquitous quantity—in physics, mathematics, and even biology—called *entropy*; and the entropy of X is 3 bits. This means that, on average, we'll have to ask 3 yes-or-no questions to find out X's value. For example, someone could tell us X's value using this code:

$$\begin{array}{llll} X_1: 001 & X_2: 010 & X_3: 011 & X_4: 100 \\ X_5: 101 & X_6: 110 & X_7: 111 & X_8: 000 \end{array}$$

So if I asked “Is the first symbol 1?” and heard “yes”, then asked “Is the second symbol 1?” and heard “no”, then asked “Is the third symbol 1?” and heard “no”, I would know that X was in state 4.

Now suppose that the system Y has four possible states with the following probabilities:

$$\begin{array}{llll} Y_1: 1/2 & Y_2: 1/4 & Y_3: 1/8 & Y_4: 1/8 \\ (50\%) & (25\%) & (12.5\%) & (12.5\%) \end{array}$$

Then the entropy of Y would be 1.75 bits, meaning that we can find out its value by asking 1.75 yes-or-no questions.

What does it mean to talk about asking one and three-fourths of a question? Imagine that we designate the states of Y using the following code:

$$Y_1: 1 \quad Y_2: 01 \quad Y_3: 001 \quad Y_4: 000$$

---

1. [http://lesswrong.com/lw/01/entropy\\_and\\_short\\_codes/](http://lesswrong.com/lw/01/entropy_and_short_codes/)

2. Page 267, ‘Where to Draw the Boundary?’.

First you ask, “Is the first symbol 1?” If the answer is “yes”, you’re done: Y is in state 1. This happens half the time, so 50% of the time, it takes 1 yes-or-no question to find out Y’s state.

Suppose that instead the answer is “No”. Then you ask, “Is the second symbol 1?” If the answer is “yes”, you’re done: Y is in state 2. Y is in state 2 with probability 1/4, and each time Y is in state 2 we discover this fact using two yes-or-no questions, so 25% of the time it takes 2 questions to discover Y’s state.

If the answer is “No” twice in a row, you ask “Is the third symbol 1?” If “yes”, you’re done and Y is in state 3; if “no”, you’re done and Y is in state 4. The 1/8 of the time that Y is in state 3, it takes three questions; and the 1/8 of the time that Y is in state 4, it takes three questions.

$$\begin{aligned} & (1/2 * 1) + (1/4 * 2) + (1/8 * 3) + (1/8 * 3) \\ &= 0.5 + 0.5 + 0.375 + 0.375 \\ &= 1.75. \end{aligned}$$

The general formula for the entropy of a system S is the sum, over all  $S_i$ , of  $-p(S_i) \log_2(p(S_i))$ .

For example, the  $\log$  (base 2) of 1/8 is -3. So  $-(1/8 * -3) = 0.375$  is the contribution of state  $S_4$  to the total entropy: 1/8 of the time, we have to ask 3 questions.

You can’t always devise a perfect code for a system, but if you have to tell someone the state of arbitrarily many copies of S in a single message, you can get arbitrarily close to a perfect code. (Google “arithmetic coding” for a simple method.)

Now, you might ask: “Why not use the code 10 for  $Y_4$ , instead of 000? Wouldn’t that let us transmit messages more quickly?”

But if you use the code 10 for  $Y_4$ , then when someone answers “Yes” to the question “Is the first symbol 1?”, you won’t know yet whether the system state is  $Y_1$  (1) or  $Y_4$  (10). In fact, if you change the code this way, the whole system falls apart—because if you hear “1001”, you don’t know if it means “ $Y_4$ , followed by  $Y_2$ ” or “ $Y_1$ , followed by  $Y_3$ .“

The moral is that *short words are a conserved resource*.

The key to creating a good code—a code that transmits messages as compactly as possible—is to reserve short words for things

that you'll need to say frequently, and use longer words for things that you won't need to say as often.

When you take this art to its limit, the length of the message you need to describe something, corresponds exactly or almost exactly to its probability. This is the Minimum Description Length or Minimum Message Length formalization of Occam's Razor<sup>3</sup>.

And so even the *labels* that we use for words are not quite arbitrary. The sounds that we attach to our concepts can be better or worse, wiser or more foolish. Even apart from considerations of common usage<sup>4</sup>!

I say all this, because the idea that “You can X any way you like” is a huge obstacle to learning how to X wisely. “It’s a free country; I have a right to my own opinions<sup>5</sup>” obstructs the art of finding truth. “I can define a word any way I like” obstructs the art of carving reality at its joints<sup>6</sup>. And even the sensible-sounding “The labels we attach to words are arbitrary” obstructs awareness of compactness. Prosody too, for that matter—Tolkien once observed what a beautiful sound the phrase “cellar door” makes; that is the kind of awareness it takes to use language like Tolkien.

The length of words also plays a nontrivial role in the cognitive science of language:

Consider the phrases “recliner”, “chair”, and “furniture”. Recliner is a more specific category than chair; furniture is a more general category than chair. But the vast majority of chairs have a common use—you use the same sort of motor actions to sit down in them, and you sit down in them for the same sort of purpose (to take your weight off your feet while you eat, or read, or type, or rest). Recliners do not depart from this theme. “Furniture”, on the other hand, includes things like beds and tables which have different uses, and call up different motor functions, from chairs.

In the terminology of cognitive psychology, “chair” is a *basic-level category*.

People have a tendency to talk, and presumably think, at the basic level of categorization—to draw the boundary around “chairs”,

---

3. Page 95, ‘Occam’s Razor’.

4. Page 238, ‘The Argument from Common Usage’.

5. [http://www.overcomingbias.com/2006/12/you\\_are\\_never\\_e.html](http://www.overcomingbias.com/2006/12/you_are_never_e.html)

6. Page 267, ‘Where to Draw the Boundary?’.

rather than around the more specific category “recliner”, or the more general category “furniture”. People are more likely to say “You can sit in that chair” than “You can sit in that recliner” or “You can sit in that furniture”.

And it is no coincidence that the word for “chair” contains fewer syllables than either “recliner” or “furniture”. Basic-level categories, in general, tend to have short names; and nouns with short names tend to refer to basic-level categories. Not a perfect rule, of course, but a definite tendency. Frequent use goes along with short words; short words go along with frequent use.

Or as Douglas Hofstadter put it, there's a reason why the English language uses “the” to mean “the” and “antidisestablishmentarianism” to mean “antidisestablishmentarianism” instead of antidisestablishmentarianism other way around.

## 23. Mutual Information, and Density in Thingspace<sup>1</sup>

### Continuation of: Entropy, and Short Codes<sup>2</sup>

Suppose you have a system X that can be in any of 8 states, which are all equally probable (relative to your current state of knowledge), and a system Y that can be in any of 4 states, all equally probable.

The entropy of X, as defined yesterday, is 3 bits; we'll need to ask 3 yes-or-no questions to find out X's exact state. The entropy of Y, as defined yesterday, is 2 bits; we have to ask 2 yes-or-no questions to find out Y's exact state. This may seem obvious since  $2^3 = 8$  and  $2^2 = 4$ , so 3 questions can distinguish 8 possibilities and 2 questions can distinguish 4 possibilities; but remember that if the possibilities were not all equally likely, we could use a more clever code to discover Y's state using e.g. 1.75 questions on average. In this case, though, X's *probability mass* is *evenly distributed* over all its possible states, and likewise Y, so we can't use any clever codes.

What is the entropy of the combined system (X,Y)?

You might be tempted to answer, "It takes 3 questions to find out X, and then 2 questions to find out Y, so it takes 5 questions total to find out the state of X and Y."

But what if the two variables are entangled, so that learning the state of Y tells us something about the state of X?

In particular, let's suppose that X and Y are either both odd, or both even.

Now if we receive a 3-bit message (ask 3 questions) and learn that X is in state 5, we know that Y is in state 1 or state 3, but not state 2 or state 4. So the single additional question "Is Y in state 3?", answered "No", tells us the entire state of (X,Y):  $X=X_5$ ,  $Y=Y_1$ . And we learned this with a total of 4 questions.

Conversely, if we learn that Y is in state 4 using two questions, it will take us only an additional two questions to learn whether X is in state 2, 4, 6, or 8. Again, four questions to learn the state of the joint system.

---

1. [http://lesswrong.com/lw/02/mutual\\_information\\_and\\_density\\_in\\_thingspace/](http://lesswrong.com/lw/02/mutual_information_and_density_in_thingspace/)

2. Page 271, 'Entropy, and Short Codes'.

The *mutual information* of two variables is defined as the difference between the entropy of the joint system and the entropy of the independent systems:  $I(X;Y) = H(X) + H(Y) - H(X,Y)$ .

Here there is one bit of mutual information between the two systems: Learning X tells us one bit of information about Y (cuts down the space of possibilities from 4 to 2, a factor-of-2 decrease in the volume) and learning Y tells us one bit of information about X (cuts down the possibility space from 8 to 4).

What about when probability mass is not evenly distributed? Yesterday, for example, we discussed the case in which Y had the probabilities  $1/2, 1/4, 1/8, 1/8$  for its four states. Let us take this to be our probability distribution over Y, considered independently - if we saw Y, without seeing anything else, this is what we'd expect to see. And suppose the variable Z has two states, 1 and 2, with probabilities  $3/8$  and  $5/8$  respectively.

Then if and only if the joint distribution of Y and Z is as follows, there is zero mutual information between Y and Z:

$$\begin{array}{llll} Z_1 Y_1: 3/16 & Z_1 Y_2: 3/32 & Z_1 Y_3: 3/64 & Z_1 Y_4: 3/64 \\ Z_2 Y_1: 5/16 & Z_2 Y_2: 5/32 & Z_2 Y_3: 5/64 & Z_2 Y_4: 5/64 \end{array}$$

This distribution obeys the law:

$$p(Y,Z) = P(Y)P(Z)$$

For example,  $P(Z_1 Y_2) = P(Z_1)P(Y_2) = 3/8 * 1/4 = 3/32$ .

And observe that we can recover the marginal (independent) probabilities of Y and Z just by looking at the joint distribution:

$$\begin{aligned} P(Y_1) &= \text{total probability of all the different ways } Y_1 \text{ can happen} \\ &= P(Z_1 Y_1) + P(Z_2 Y_1) \\ &= 3/16 + 5/16 \\ &= 1/2. \end{aligned}$$

So, just by inspecting the joint distribution, we can determine whether the marginal variables Y and Z are independent; that is,

whether the joint distribution factors into the product of the marginal distributions; whether, for all Y and Z,  $P(Y,Z) = P(Y)P(Z)$ .

This last is significant because, by Bayes's Rule<sup>3</sup>:

$$\begin{aligned} P(Y_i, Z_j) &= P(Y_i)P(Z_j) \\ P(Y_i, Z_j)/P(Z_j) &= P(Y_i) \\ P(Y_i|Z_j) &= P(Y_i) \end{aligned}$$

In English, "After you learn  $Z_j$ , your belief about  $Y_i$  is just what it was before."

So when the distribution factorizes - when  $P(Y,Z) = P(Y)P(Z)$  - this is *equivalent* to "Learning about Y never tells us anything about Z or vice versa."

From which you might suspect, correctly, that there is no mutual information between Y and Z. Where there is no mutual information, there is no Bayesian evidence, and vice versa.

Suppose that in the distribution YZ above, we treated each possible combination of Y and Z as a separate event—so that the distribution YZ would have a total of 8 possibilities, with the probabilities shown—and then we calculated the entropy of the distribution YZ the same way we would calculate the entropy of any distribution:

$$3/16 \log_2(3/16) + 3/32 \log_2(3/32) + 3/64 \log_2(3/64) + \dots + 5/64 \log_2(5/64)$$

You would end up with the same total you would get if you separately calculated the entropy of Y plus the entropy of Z. There is no mutual information between the two variables, so our uncertainty about the joint system is not any less than our uncertainty about the two systems considered separately. (I am not showing the calculations, but you are welcome to do them; and I am not showing the proof that this is true in general, but you are welcome to Google on "Shannon entropy" and "mutual information".)

What if the joint distribution doesn't factorize? For example:

$$Z_1 Y_1: 12/64 \quad Z_1 Y_2: 8/64 \quad Z_1 Y_3: 1/64 \quad Z_1 Y_4: 3/64$$

---

<sup>3</sup>. Page 28, 'An Intuitive Explanation of Bayes' Theorem'.

$$Z_2 Y_1: 20/64 \quad Z_2 Y_2: 8/64 \quad Z_2 Y_3: 7/64 \quad Z_2 Y_4: 5/64$$

If you add up the joint probabilities to get marginal probabilities, you should find that  $P(Y_1) = 1/2$ ,  $P(Z_1) = 3/8$ , and so on - the marginal probabilities are the same as before.

But the joint probabilities do not always equal the product of the marginal probabilities. For example, the probability  $P(Z_1 Y_2)$  equals  $8/64$ , where  $P(Z_1)P(Y_2)$  would equal  $3/8 * 1/4 = 6/64$ . That is, the probability of running into  $Z_1 Y_2$  together, is greater than you'd expect based on the probabilities of running into  $Z_1$  or  $Y_2$  separately.

Which in turn implies:

$$\begin{aligned} P(Z_1 Y_2) &> P(Z_1)P(Y_2) \\ P(Z_1 Y_2)/P(Y_2) &> P(Z_1) \\ P(Z_1|Y_2) &> P(Z_1) \end{aligned}$$

Since there's an "unusually high" probability for  $P(Z_1 Y_2)$  - defined as a probability higher than the marginal probabilities would indicate by default - it follows that observing  $Y_2$  is evidence which increases the probability of  $Z_1$ . And by a symmetrical argument, observing  $Z_1$  must favor  $Y_2$ .

As there are at least some values of  $Y$  that tell us about  $Z$  (and vice versa) there must be mutual information between the two variables; and so you will find—I am confident, though I haven't actually checked—that calculating the entropy of  $YZ$  yields less total uncertainty than the sum of the independent entropies of  $Y$  and  $Z$ .  $H(Y,Z) = H(Y) + H(Z) - I(Y;Z)$  with all quantities necessarily non-negative.

(I digress here to remark that the symmetry of the expression for the mutual information shows that  $Y$  *must* tell us as much about  $Z$ , on average, as  $Z$  tells us about  $Y$ . I leave it as an exercise to the reader to reconcile this with anything they were taught in logic class about how, if all ravens are black, being allowed to reason  $\text{Raven}(x) \rightarrow \text{Black}(x)$  doesn't mean you're allowed to reason  $\text{Black}(x) \rightarrow \text{Raven}(x)$ . How different seem the symmetrical

probability flows of the Bayesian, from the sharp lurches of logic—even though the latter is just a degenerate case of the former.)

“But,” you ask, “what has all this to do with the proper use of words?”

In Empty Labels<sup>4</sup> and then Replace the Symbol with the Substance<sup>5</sup>, we saw the technique of replacing a word with its definition – the example being given:

All [mortal, -feathers, bipedal] are mortal.  
 Socrates is a [mortal, -feathers, bipedal].  
 Therefore, Socrates is mortal.

Why, then, would you even want to have a word for “human”? Why not just say “Socrates is a mortal featherless biped”?

Because it’s helpful to have shorter words for things that you encounter often. If your code for describing single properties is already efficient, then there will not be an advantage to having a special word for a conjunction – like “human” for “mortal featherless biped” – unless things that are mortal *and* featherless *and* bipedal, are found *more often* than the marginal probabilities would lead you to expect.

In efficient codes, word length corresponds to probability—so the code for  $Z_1 Y_2$  will be just as long as the code for  $Z_1$  plus the code for  $Y_2$ , unless  $P(Z_1 Y_2) > P(Z_1)P(Y_2)$ , in which case the code for the word can be shorter than the codes for its parts.

And this in turn corresponds exactly to the case where we can infer some of the properties of the thing, from seeing its other properties. It must be more likely than the default that featherless bipedal things will also be mortal.

Of course the word “human” really describes many, many more properties – when you see a human-shaped entity that talks and wears clothes, you can infer whole hosts of biochemical and anatomical and cognitive facts about it. To replace the word “human” with a description of everything we know about humans would require us to spend an inordinate amount of time talking.

---

4. Page 242, ‘Empty Labels’.

5. Page 248, ‘Replace the Symbol with the Substance’.

But this is true *only* because a featherless talking biped is far more likely than default to be poisonable by hemlock, or have broad nails, or be overconfident.

Having a word for a thing, rather than just listing its properties, is a more compact code precisely in those cases where we can infer some of those properties from the other properties. (With the exception perhaps of very primitive words, like “red”, that we would use to send an entirely uncompressed description of our sensory experiences. But by the time you encounter a bug, or even a rock, you’re dealing with nonsimple property collections, far above the primitive level.)

So having a word “wiggin”<sup>6</sup> for green-eyed black-haired people, is more useful than just saying “green-eyed black-haired person”, precisely when:

1. Green-eyed people are more likely than average to be black-haired (and vice versa), meaning that we can probabilistically infer green eyes from black hair or vice versa; *or*
2. Wiggins share other properties that can be inferred at greater-than-default probability. In this case we have to separately observe the green eyes and black hair; but then, after observing both these properties independently, we can probabilistically infer other properties (like a taste for ketchup).

One may even consider the act of defining a word as a promise to this effect. Telling someone, “I define the word ‘wiggin’ to mean a person with green eyes and black hair”, by Gricean implication, asserts that the word “wiggin” will somehow help you make inferences / shorten your messages.

If green-eyes and black hair have no greater than default probability to be found together, nor does any other property occur at greater than default probability along with them, then the word “wiggin” is a lie: The word claims that certain people are worth distinguishing as a group, but they’re not.

In this case the word “wiggin” does not help describe reality more compactly—it is not defined by someone sending the shortest message—it has no role in the simplest explanation. Equivalently,

---

6. Page 260, ‘Sneaking in Connotations’.

the word “wiggin” will be of no help to you in doing any Bayesian inference. Even if you do not call the word a lie, it is surely an error.

And the way to carve reality at its joints, is to draw your boundaries around concentrations of unusually high probability density in Thingspace<sup>7</sup>.

---

7. Page 209, ‘The Cluster Structure of Thingspace’.

## 24. Superexponential Conceptspace, and Simple Words<sup>1</sup>

**Followup to:** Mutual Information, and Density in Thingspace<sup>2</sup>

Thingspace<sup>3</sup>, you might think, is a rather huge space. Much larger than reality, for where reality only contains things that actually exist, Thingspace contains everything that *could* exist.

Actually, the way I “defined” Thingspace to have dimensions for every possible attribute—including correlated attributes like density and volume and mass—Thingspace may be too poorly defined to have anything you could call a *size*. But it’s important to be able to visualize Thingspace *anyway*. Surely, no one can *really* understand a flock of sparrows if all they see is a cloud of flapping cawing things, rather than a cluster of points in Thingspace.

But as vast as Thingspace may be, it doesn’t hold a candle to the size of Conceptspace.

“Concept”, in machine learning, means a rule that includes or excludes examples. If you see the data 2:+, 3:-, 14:+, 23:-, 8:+, 9:- then you might guess that the concept was “even numbers”. There is a rather large literature (as one might expect) on how to learn concepts from data... given random examples, given chosen examples... given possible errors in classification... and most importantly, given different spaces of possible rules.

Suppose, for example, that we want to learn the concept “good days on which to play tennis”. The possible attributes of Days are:

Sky:	{Sunny, Cloudy, Rainy}
AirTemp:	{Warm, Cold}
Humidity:	{Normal, High}
Wind:	{Strong, Weak}

We’re then presented with the following data, where + indicates a positive example of the concept, and - indicates a negative classification:

---

1. [http://lesswrong.com/lw/03/superexponential\\_conceptspace\\_and\\_simple\\_words/](http://lesswrong.com/lw/03/superexponential_conceptspace_and_simple_words/)

2. Page 275, ‘Mutual Information, and Density in Thingspace’.

3. Page 209, ‘The Cluster Structure of Thingspace’.

+	Sky: Sunny;	AirTemp: Warm;
	Humidity: High;	Wind: Strong.
-	Sky: Rainy;	AirTemp: Cold;
	Humidity: High;	Wind: Strong.
+	Sky: Sunny;	AirTemp: Warm;
	Humidity: High;	Wind: Weak.

What should an algorithm infer from this?

A machine learner might represent *one* concept that fits this data as follows:

```
Sky: ?; AirTemp: Warm; Humidity:  
High; Wind: ?
```

In this format, to determine whether this concept accepts or rejects an example, we compare element-by-element: `?` accepts anything, but a specific value accepts only that specific value.

So the concept above will accept only Days with  $\text{AirTemp}=\text{Warm}$  and  $\text{Humidity}=\text{High}$ , but the Sky and the Wind can take on any value. This fits both the negative and the positive classifications in the data so far—though it isn’t the *only* concept that does so.

We can also simplify the above concept representation to `{?, Warm, High, ?}`.

Without going into details, the classic algorithm would be:

- Maintain the set of the most general hypotheses that fit the data—those that positively classify as many examples as possible, while still fitting the facts.
- Maintain another set of the most specific hypotheses that fit the data—those that negatively classify as many examples as possible, while still fitting the facts.
- Each time we see a new negative example, we strengthen all the most general hypotheses as little as possible, so that the new set is again as general as possible while fitting the facts.
- Each time we see a new positive example, we relax all the most specific hypotheses as little as possible, so that the

new set is again as specific as possible while fitting the facts.

- We continue until we have only a single hypothesis left. This will be the answer if the target concept was in our hypothesis space at all.

In the case above, the set of most general hypotheses would be  $\{?, \text{Warm}, ?, ?, ?\}$  and  $\{\text{Sunny}, ?, ?, ?, ?\}$ , while the set of most specific hypotheses is the single member  $\{\text{Sunny, Warm, High, ?}\}$ .

Any other concept you can find that fits the data will be strictly more specific than one of the most general hypotheses, and strictly more general than the most specific hypothesis.

(For more on this, I recommend Tom Mitchell's *Machine Learning*, from which this example was adapted.)

Now you may notice that the format above *cannot* represent all possible concepts. E.g. "Play tennis when the sky is sunny *or* the air is warm". That fits the data, but in the concept representation defined above, there's no quadruplet of values that describes the rule.

Clearly our machine learner is not very general. Why not allow it to represent *all possible* concepts, so that it can learn with the greatest possible flexibility?

Days are composed of these four variables, one variable with 3 values and three variables with 2 values. So there are  $3^2 \cdot 2^2 = 24$  possible Days that we could encounter.

The format given for representing Concepts allows us to require any of these values for a variable, or leave the variable open. So there are  $4^*3^*3^*3 = 108$  concepts in that representation. For the most-general/most-specific algorithm to work, we need to start with the most specific hypothesis "no example is ever positively classified". If we add that, it makes a total of 109 concepts.

Is it suspicious that there are more possible concepts than possible Days? Surely not: After all, a concept can be viewed as a *collection* of Days. A concept can be viewed as the set of days that it classifies positively, or isomorphically, the set of days that it classifies negatively.

So the space of *all possible* concepts that classify Days is the set of all possible sets of Days, whose size is  $2^{24} = 16,777,216$ .

This complete space includes all the concepts we have discussed so far. But it also includes concepts like "Positively classify only the

examples {Sunny, Warm, High, Strong} and {Sunny, Warm, High, Weak} and reject everything else” or “Negatively classify only the example {Rainy, Cold, High, Strong} and accept everything else.” It includes concepts with no compact representation, just a flat list of what is and isn’t allowed.

That’s the problem with trying to build a “fully general” inductive learner: They can’t learn concepts until they’ve seen every possible example in the instance space.

If we add on more attributes to Days—like the Water temperature, or the Forecast for tomorrow—then the number of possible days will grow exponentially in the number of attributes. But this isn’t a problem with our restricted concept space, because you can narrow down a large space using a logarithmic number of examples.

Let’s say we add the Water: {Warm, Cold} attribute to days, which will make for 48 possible Days and 325 possible concepts. Let’s say that each Day we see is, usually, classified positive by around half of the currently-plausible concepts, and classified negative by the other half. Then when we learn the actual classification of the example, it will cut the space of compatible concepts in half. So it might only take 9 examples ( $2^9 = 512$ ) to narrow 325 possible concepts down to one.

Even if Days had forty binary attributes, it should still only take a manageable amount of data to narrow down the possible concepts to one. 64 examples, if each example is classified positive by half the remaining concepts. *Assuming*, of course, that the *actual* rule is one we can represent at all!

If you want to think of all the possibilities, well, good luck with that. The space of *all possible* concepts grows *superexponentially* in the number of attributes.

By the time you’re talking about data with forty binary attributes, the number of possible examples is past a trillion—but the number of possible *concepts* is past two-to-the-trillionth-power. To narrow down that *superexponential* concept space, you’d have to see over a trillion examples before you could say what was In, and what was Out. You’d have to see every possible example, in fact.

That’s with forty binary attributes, mind you. 40 bits, or 5 bytes, to be classified simply “Yes” or “No”. 40 bits implies  $2^{40}$

possible examples, and  $2^{(2^40)}$  possible concepts that classify those examples as positive or negative.

So, here in the real world, where objects take more than 5 bytes to describe *and* a trillion examples are not available *and* there is noise in the training data, we only even *think* about *highly regular* concepts. A human mind—or the whole observable universe—is not nearly large enough to consider all the other hypotheses.

From this perspective, learning doesn't just *rely on* inductive bias<sup>4</sup>, it is *nearly all* inductive bias—when you compare the number of concepts ruled out *a priori*, to those ruled out by mere evidence.

But what has this (you inquire) to do with the proper use of words?

It's the whole reason that words have intensions as well as extensions<sup>5</sup>.

In yesterday's post<sup>6</sup>, I concluded:

The way to carve reality at its joints, is to draw boundaries around concentrations of unusually high probability density.

I deliberately left out a key qualification in that (slightly edited) statement, because I couldn't explain it until today. A better statement would be:

The way to carve reality at its joints, is to draw *simple* boundaries around concentrations of unusually high probability density in Thingspace.

Otherwise you would just gerrymander Thingspace. You would create really odd noncontiguous boundaries that collected the observed examples, examples that couldn't be described in any shorter message<sup>7</sup> than your observations themselves, and say: "This is what I've seen before, and what I expect to see more of in the future."

In the real world, nothing above the level of molecules repeats itself *exactly*. Socrates is shaped a lot like all those other humans

4. [http://lesswrong.com/lw/hg/inductive\\_bias/](http://lesswrong.com/lw/hg/inductive_bias/)

5. Page 200, 'Extensions and Intensions'.

6. Page 275, 'Mutual Information, and Density in Thingspace'.

7. Page 271, 'Entropy, and Short Codes'.

who were vulnerable to hemlock, but he isn't shaped *exactly* like them. So your guess that Socrates is a "human" relies on drawing *simple* boundaries around the human cluster in Thingspace. Rather than, "Things shaped exactly like [5-megabyte shape specification 1] and with [lots of other characteristics], *or* exactly like [5-megabyte shape specification 2] and [lots of other characteristics]", ..., are human."

If you don't draw *simple* boundaries around your experiences, you can't do inference with them. So you try to describe "art"<sup>8</sup> with intensional definitions like "that which is intended to inspire any complex emotion for the sake of inspiring it", rather than just pointing at a long list of things that are, or aren't art.

In fact, the above statement about "how to carve reality at its joints" is a bit chicken-and-eggish: You can't assess the *density* of actual observations, until you've already done at least a little carving. And the probability distribution comes from drawing the boundaries, not the other way around—if you already *had* the probability distribution, you'd have everything necessary for inference, so why would you bother drawing boundaries?

And this suggests another—yes, yet another—reason to be suspicious of the claim that "you can define a word any way you like". When you consider the superexponential size of Conceptspace, it becomes clear that singling out one particular concept for consideration<sup>9</sup> is an act of no small audacity—not just for us, but for any mind of bounded computing power.

Presenting us with the word "wiggin", defined as "a black-haired green-eyed person", without some reason for raising *this particular concept* to the level of our deliberate attention, is rather like a detective saying: "Well, I haven't the slightest shred of support one way or the other for who could've murdered those orphans... not even an intuition, mind you... but have we considered John Q. Wiffleheim of 1234 Norkle Rd as a suspect?"

---

8. Page 267, 'Where to Draw the Boundary?'.

9. [http://lesswrong.com/lw/jo/einsteins\\_arrogance/](http://lesswrong.com/lw/jo/einsteins_arrogance/)

## 25. Conditional Independence, and Naive Bayes<sup>1</sup>

**Followup to:** Searching for Bayes-Structure<sup>2</sup>

Previously I spoke of mutual information<sup>3</sup> between X and Y,  $I(X;Y)$ , which is the difference between the **entropy**<sup>4</sup> of the joint probability distribution,  $H(X,Y)$  and the entropies of the marginal distributions,  $H(X) + H(Y)$ .

I gave the example of a variable X, having eight states 1..8 which are all equally probable if we have not yet encountered any evidence; and a variable Y, with states 1..4, which are all equally probable if we have not yet encountered any evidence. Then if we calculate the marginal entropies  $H(X)$  and  $H(Y)$ , we will find that X has 3 bits of entropy, and Y has 2 bits.

However, we also know that X and Y are both even or both odd; and this is all we know about the relation between them. So for the joint distribution  $(X,Y)$  there are only 16 possible states, all equally probable, for a joint entropy of 4 bits. This is a 1-bit entropy defect, compared to 5 bits of entropy if X and Y were independent. This entropy defect is the mutual information - the information that X tells us about Y, or vice versa, so that we are not as uncertain about one after having learned the other.

Suppose, however, that there exists a third variable Z. Z has two states, “even” and “odd”, perfectly correlated to the evenness or oddness of  $(X,Y)$ . In fact, we’ll suppose that Z is just the question “Are X and Y even or odd?”

If we have no evidence about X and Y, then Z itself necessarily has 1 bit of entropy on the information given. There is 1 bit of mutual information between Z and X, and 1 bit of mutual information between Z and Y. And, as previously noted, 1 bit of mutual information between X and Y. So how much entropy for the whole system  $(X,Y,Z)$ ? You might naively expect that

---

1. [http://lesswrong.com/lw/08/conditional\\_independence\\_and\\_naive\\_bayes/](http://lesswrong.com/lw/08/conditional_independence_and_naive_bayes/)

2. [http://lesswrong.com/lw/07/searching\\_for\\_bayesstructure/](http://lesswrong.com/lw/07/searching_for_bayesstructure/)

3. Page 275, ‘Mutual Information, and Density in Thingspace’.

4. Page 271, ‘Entropy, and Short Codes’.

$$H(X,Y,Z) = H(X) + H(Y) + H(Z) - I(X;Z) - I(Z;Y) - I(X;Y)$$

but this turns out not to be the case.

The joint system  $(X,Y,Z)$  only has 16 possible states - since  $Z$  is just the question “Are  $X$  &  $Y$  even or odd?” - so  $H(X,Y,Z) = 4$  bits.

But if you calculate the formula just given, you get

$$(3 + 2 + 1 - 1 - 1 - 1)\text{bits} = 3 \text{ bits} = \text{WRONG!}^5$$

Why? Because if you have the mutual information between  $X$  and  $Z$ , and the mutual information between  $Z$  and  $Y$ , that may include some of the *same* mutual information that we'll calculate exists between  $X$  and  $Y$ . In this case, for example, knowing that  $X$  is even tells us that  $Z$  is even, and knowing that  $Z$  is even tells us that  $Y$  is even, but this is the same information that  $X$  would tell us about  $Y$ . We double-counted<sup>6</sup> some of our knowledge, and so came up with too little entropy.

The correct formula is (I believe):

$$H(X,Y,Z) = H(X) + H(Y) + H(Z) - I(X;Z) - I(Z;Y) - I(X;Y | Z)$$

Here the last term,  $I(X;Y | Z)$ , means, “the information that  $X$  tells us about  $Y$ , given that we already know  $Z$ ”. In this case,  $X$  doesn't tell us anything about  $Y$ , given that we already know  $Z$ , so the term comes out as zero - and the equation gives the correct answer. There, isn't that nice?

“No,” you correctly<sup>7</sup> reply, “for you have not told me how to *calculate*  $I(X;Y|Z)$ , only given me a verbal argument that it ought to be zero.”

We calculate  $I(X;Y|Z)$  just the way you would expect.  $I(X;Y) = H(X) + H(Y) - H(X,Y)$ , so:

$$I(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z)$$

5. <http://www.youtube.com/watch?v=tRVUOGUmxfI>

6. Page 147, ‘Fake Causality’.

7. <http://yudkowsky.net/bayes/technical.html>

And now, I suppose, you want to know how to calculate the conditional entropy? Well, the *original* formula for the entropy is:

$$H(S) = \text{Sum } i: p(S_i)^* \cdot \log_2(p(S_i))$$

If we then learned a new fact  $Z_o$ , our remaining uncertainty about S would be:

$$H(S|Z_o) = \text{Sum } i: p(S_i|Z_o)^* \cdot \log_2(p(S_i|Z_o))$$

So if we're going to learn a new fact Z, but we don't know which Z yet, then, on average, we expect to be around this uncertain of S afterward:

$$H(S|Z) = \text{Sum } j: (p(Z_j) * \text{Sum } i: p(S_i|Z_j)^* \cdot \log_2(p(S_i|Z_j)))$$

And that's how one calculates conditional entropies; from which, in turn, we can get the conditional mutual information.

There are *all sorts* of ancillary theorems here, like:

$$H(X|Y) = H(X, Y) - H(Y)$$

and

$$\text{if } I(X;Z) = 0 \text{ and } I(Y;X|Z) = 0 \text{ then } I(X;Y) = 0$$

but I'm not going to go into those.

"But," you ask, "what does *this* have to do with the nature of words and their hidden Bayesian structure?"

I am just so *unspeakably* glad that you asked that question, because I was planning to tell you whether you liked it or not. But first there are a couple more preliminaries.

You will remember—yes, you *will* remember—that there is a duality between mutual information and Bayesian evidence. Mutual information is positive if and only if the probability of at least some joint events  $P(x, y)$  does not equal the product of the probabilities of the separate events  $P(x)*P(y)$ . This, in turn, is exactly equivalent to the condition that Bayesian evidence exists between x and y:

$$\begin{aligned} I(X;Y) > 0 &\Rightarrow \\ P(x,y) &\neq P(x)*P(y) \\ P(x,y) / P(y) &\neq P(x) \\ P(x|y) &\neq P(x) \end{aligned}$$

If you're conditioning on  $Z$ , you just adjust the whole derivation accordingly:

$$\begin{aligned} I(X;Y | Z) > 0 &\Rightarrow \\ P(x,y|z) &\neq P(x|z)*P(y|z) \\ P(x,y|z) / P(y|z) &\neq P(x|z) \\ (P(x,y,z) / P(z)) / (P(y,z) / P(z)) &\neq P(x|z) \\ P(x,y,z) / P(y,z) &\neq P(x|z) \\ P(x|y,z) &\neq P(x|z) \end{aligned}$$

Which last line reads “Even knowing  $Z$ , learning  $Y$  still changes our beliefs about  $X$ .”

Conversely, as in our original case of  $Z$  being “even” or “odd”,  $Z$  screens off<sup>8</sup>  $X$  from  $Y$  - that is, if we know that  $Z$  is “even”, learning that  $Y$  is in state 4 tells us *nothing more* about whether  $X$  is 2, 4, 6, or 8. Or if we know that  $Z$  is “odd”, then learning that  $X$  is 5 tells us nothing more about whether  $Y$  is 1 or 3. Learning  $Z$  has rendered  $X$  and  $Y$  *conditionally independent*.

Conditional independence is a hugely important concept in probability theory—to cite just one example, without conditional independence, the universe would have no structure.

Today, though, I only intend to talk about one particular kind of conditional independence—the case of a central variable that screens off other variables surrounding it, like a central body with tentacles.

Let there be five variables  $U, V, W, X, Y$ ; and moreover, suppose that for every pair of these variables, one variable is evidence about the other. If you select  $U$  and  $W$ , for example, then learning  $U=U_1$  will tell you something you didn't know before about the probability  $W=W_1$ .

An unmanageable inferential mess? Evidence gone wild? Not necessarily.

---

8. Page 338, ‘Argument Screens Off Authority’.

Maybe U is “Speaks a language”, V is “Two arms and ten digits”, W is “Wears clothes”, X is “Poisonable by hemlock”, and Y is “Red blood”. Now if you encounter a thing-in-the-world, that might be an apple and might be a rock, and you learn that this thing speaks Chinese, you are liable to assess a much higher probability that it wears clothes; and if you learn that the thing is not poisonable by hemlock, you will assess a somewhat lower probability that it has red blood.

Now some of these rules are stronger than others. There is the case of Fred, who is missing a finger due to a volcano accident, and the case of Barney the Baby who doesn’t speak yet, and the case of Irving the IRCBot who emits sentences but has no blood. So if we learn that a certain thing is not wearing clothes, that doesn’t screen off everything that its speech capability can tell us about its blood color. If the thing doesn’t wear clothes but *does* talk, maybe it’s Nude Nellie.

This makes the case more interesting than, say, five integer variables that are all odd or all even, but otherwise uncorrelated. In that case, knowing *any* one of the variables would screen off everything that knowing a second variable could tell us about a third variable.

But here, we have dependencies that don’t go away as soon as we learn just one variable, as the case of Nude Nellie shows. So is it an unmanageable inferential inconvenience?

Fear not! for there may be some *sixth* variable Z, which, if we knew it, really *would* screen off every pair of variables from each other. There may be some variable Z—even if we have to *construct* Z rather than observing it directly—such that:

$$\begin{aligned} p(u|v,w,x,y,z) &= p(u|z) \\ p(v|u,w,x,y,z) &= p(v|z) \\ p(w|u,v,x,y,z) &= p(w|z) \end{aligned}$$

...

Perhaps, *given that* a thing is “human”, then the probabilities of it speaking, wearing clothes, and having the standard number of fingers, are all independent. Fred may be missing a finger - but he is no more likely to be a nudist than the next person; Nude Nellie never wears clothes, but knowing this doesn’t make it any less likely that

she speaks; and Baby Barney doesn't talk yet, but is not missing any limbs.

This is called the "Naive Bayes" method, because it usually isn't quite true, but *pretending* that it's true can simplify the living daylights out of your calculations. We don't keep separate track of the influence of clothed-ness on speech capability given finger number. We just use all the information we've observed to keep track of the probability that this thingy is a human (or alternatively, something else, like a chimpanzee or robot) and then use our beliefs about the central class to predict anything we haven't seen yet, like vulnerability to hemlock.

Any observations of U, V, W, X, and Y just act as evidence for the central class variable Z, and then we use the posterior distribution on Z to make any predictions that need making about unobserved variables in U, V, W, X, and Y.

Sound familiar? It should:

Blegg2

As a matter of fact, if you use the right kind of neural network units, this "neural network" ends up *exactly, mathematically* equivalent to Naive Bayes. The central unit just needs a logistic threshold—an S-curve response—and the weights of the inputs just need to match the logarithms of the likelihood ratios, etcetera. In fact, it's a good guess that this is one of the reasons why logistic response

often works so well in neural networks—it lets the algorithm sneak in a little Bayesian reasoning while the designers aren't looking.

Just because someone is presenting you with an algorithm that they call a “neural network” with buzzwords like “scruffy” and “emergent” plastered all over it, disclaiming proudly that they have no idea how the learned network works—well, don’t assume that their little AI algorithm *really is* Beyond the Realms of Logic. For this paradigm of adhockery, if it works, will turn out to have Bayesian structure<sup>9</sup>; it may even be exactly equivalent to an algorithm of the sort called “Bayesian”.

Even if it doesn’t *look* Bayesian, on the surface.

And then you just *know* that the Bayesians are going to start explaining exactly how the algorithm works, what underlying assumptions it reflects, which environmental regularities<sup>10</sup> it exploits, where it works and where it fails, and even attaching understandable meanings to the learned network weights.

Disappointing, isn’t it?

---

9. [http://lesswrong.com/lw/07/searching\\_for\\_bayesstructure/](http://lesswrong.com/lw/07/searching_for_bayesstructure/)

10. Page 275, ‘Mutual Information, and Density in Thingspace’.

## 26. Words as Mental Paintbrush Handles<sup>1</sup>

**Followup to:** Conditional Independence, and Naive Bayes<sup>2</sup>

(We should be done with the mathy posts, I think, at least for now. But forgive me if, ironically, I end up resorting to Rationality Quotes for a day or two. I'm currently at the AGI-08 conference, which, as of the first session, is not nearly so bad as I feared.)

Suppose I tell you: “It’s the strangest thing: The lamps in this hotel have triangular lightbulbs.”

You may or may not have visualized it—if you haven’t done it yet, do so now—what, in your mind’s eye, does a “triangular lightbulb” look like?

In your mind’s eye, did the glass have sharp edges, or smooth?

When the phrase “triangular lightbulb” first crossed my mind—no, the hotel doesn’t have them—then as best as my introspection could determine, I first saw a pyramidal lightbulb with sharp edges, then (almost immediately) the edges were smoothed, and then my mind generated a loop of florescent bulb in the shape of a smooth triangle as an alternative.

As far as I can tell, no deliberative/verbal thoughts were involved—just wordless reflex flinch away from the imaginary mental vision of sharp glass, which design problem was solved before I could even think in words.

Believe it or not, for some decades, there was a serious debate about whether people *really* had mental images in their mind—an actual *picture* of a chair somewhere—or if people just naively *thought* they had mental images (having been misled by “introspection”, a very bad forbidden activity), while actually just having a little “chair” label, like a LISP token, active in their brain.

I am trying hard not to say anything like “How spectacularly silly,” because there is always the hindsight effect<sup>3</sup> to consider, but: how spectacularly silly.

This academic paradigm, I think, was mostly a deranged legacy of behaviorism, which denied the existence of thoughts in humans, and sought to explain all human phenomena as “reflex”, including

---

1. [http://lesswrong.com/lw/09/words\\_as\\_mental\\_paintbrush\\_handles/](http://lesswrong.com/lw/09/words_as_mental_paintbrush_handles/)

2. Page 288, ‘Conditional Independence, and Naive Bayes’.

3. Page 136, ‘Hindsight Devalues Science’.

speech. Behaviorism probably deserves its own post at some point, as it was a perversion of rationalism; but this is not that post.

“You call it ‘silly’,” you inquire, “but how do you *know* that your brain represents visual images? Is it merely that you can close your eyes and see them?”

This question *used* to be harder to answer, back in the day of the controversy. If you wanted to prove the existence of mental imagery “scientifically”, rather than just by introspection, you had to infer the existence of mental imagery from experiments like, e.g.: Show subjects two objects and ask them if one can be rotated into correspondence with the other. The response time is linearly proportional to the angle of rotation required. This is easy to explain if you are actually visualizing the image and continuously rotating it at a constant speed, but hard to explain if you are just checking propositional features of the image.

Today we can actually neuroimage the little pictures in the visual cortex. So, yes, your brain really does represent a detailed image of what it sees or imagines. See Stephen Kosslyn’s *Image and Brain: The Resolution of the Imagery Debate*.

Part of the reason people get in trouble with words, is that they do not realize how much complexity lurks behind words.

Can you visualize a “green dog”? Can you visualize a “cheese apple”?

“Apple” isn’t just a sequence of two syllables or five letters. That’s a shadow. That’s the tip of the tiger’s tail.

Words, or rather the concepts behind them, are paintbrushes—you can use them to draw images in your own mind. Literally draw, if you employ concepts to make a picture in your visual cortex. And by the use of shared labels, you can reach into someone else’s mind, and grasp their paintbrushes to draw pictures in *their* minds—sketch a little green dog in their visual cortex.

But don’t think that, because you send syllables through the air, or letters through the Internet, it is the syllables or the letters that draw pictures in the visual cortex. That takes some complex instructions that wouldn’t fit in the sequence of letters. “Apple” is 5 bytes, and drawing a picture of an apple from scratch would take more data than that.

“Apple” is merely the tag attached to the true and wordless *apple* concept, which can paint a picture in your visual cortex, or collide with “cheese”, or recognize an apple when you see one, or taste its archetype in apple pie, maybe even send out the motor behavior for eating an apple...

And it’s not as simple as just calling up a picture from memory. Or how would you be able to visualize combinations like a “triangular lightbulb”—imposing triangledness on lightbulbs, keeping the essence of both, even if you’ve never seen such a thing in your life?

Don’t make the mistake the behaviorists made. There’s far more to speech than sound in air. The labels are just pointers—“look in memory area 1387540”. Sooner or later, when you’re handed a pointer, it comes time to dereference it, and actually look in memory area 1387540.

What does a word point to?

## 27. Variable Question Fallacies<sup>1</sup>

**Followup to:** Words as Mental Paintbrush Handles<sup>2</sup>

Albert: “Every time I’ve listened to a tree fall, it made a sound, so I’ll guess that other trees falling also make sounds. I don’t believe the world changes around when I’m not looking.”

Barry: “Wait a minute. If no one hears it, how can it be a sound?”

While writing the dialogue of Albert and Barry in their dispute<sup>3</sup> over whether a falling tree in a deserted forest makes a sound, I sometimes found myself losing empathy with my characters. I would start to lose the gut feel of why *anyone* would ever argue like that, even though I’d seen it happen many times.

On these occasions, I would repeat to myself, “Either the falling tree makes a sound, or it does not!” to restore my borrowed sense of indignation.

(P or  $\neg P$ ) is not always a reliable heuristic, if you substitute arbitrary English sentences for P. “This sentence is false” cannot be consistently viewed as true or false. And then there’s the old classic, “Have you stopped beating your wife?”

Now if you are a mathematician, and one who believes in classical (rather than intuitionistic) logic, there are ways to continue insisting that (P or  $\neg P$ ) is a theorem: for example, saying that “This sentence is false” is not a sentence.

But such resolutions are subtle, which suffices to demonstrate a need for subtlety. You cannot just bull ahead on every occasion with “Either it does or it doesn’t!”

So does the falling tree make a sound, or not, or...?

Surely,  $2 + 2 = X$  or it does not? Well, maybe, if it’s *really* the same X, the same 2, and the same + and =. If X evaluates to 5 on some occasions and 4 on another, your indignation may be misplaced.

---

1. [http://lesswrong.com/lw/oc/variable\\_question\\_fallacies/](http://lesswrong.com/lw/oc/variable_question_fallacies/)

2. Page 295, ‘Words as Mental Paintbrush Handles’.

3. Page 228, ‘Disputing Definitions’.

To even begin claiming that ( $P$  or  $\neg P$ ) ought to be a necessary truth, the symbol  $P$  must stand for *exactly* the same thing in both halves of the dilemma. “Either the fall makes a sound, or not!”—but if Albert::sound is not the same as Barry::sound, there is nothing paradoxical about the tree making an Albert::sound but not a Barry::sound.

(The :: idiom is something I picked up in my C++ days for avoiding namespace collisions. If you’ve got two different packages that define a class Sound, you can write Package1::Sound to specify which Sound you mean. The idiom is not widely known, I think; which is a pity, because I often wish I could use it in writing.)

The variability may be subtle: Albert and Barry may carefully verify that it is the same tree, in the same forest, and the same occasion of falling, just to ensure that they really do have a substantive disagreement about exactly the same event. And then forget to check that they are matching this event against exactly the same concept.

Think about the grocery store that you visit most often: Is it on the left side of the street, or the right? But of course there is no “*the* left side” of the street, only *your* left side, as you travel along it from some particular direction. Many of the words we use are really functions of implicit variables supplied by context.

It’s actually one heck of a pain, requiring one heck of a lot of work, to handle this kind of problem in an Artificial Intelligence program intended to parse language—the phenomenon going by the name of “speaker deixis”.

“Martin told Bob the building was on his left.” But “left” is a function-word that evaluates with a speaker-dependent variable invisibly grabbed from the surrounding context. Whose “left” is meant, Bob’s or Martin’s?

The variables in a variable question fallacy often aren’t neatly labeled—it’s not as simple as “Say, do you think  $Z + 2$  equals 6?”

If a namespace collision<sup>4</sup> introduces two different concepts that look like “the same concept” because they have the same name—or a map compression<sup>5</sup> introduces two different events that look like the same event because they don’t have separate mental files—or

4. Page 245, ‘Taboo Your Words’.

5. Page 253, ‘Fallacies of Compression’.

the same function evaluates in different contexts—then reality itself becomes protean, changeable. At least that's what the algorithm feels like from inside<sup>6</sup>. Your mind's eye sees the map, not the territory directly.

If you have a question with a hidden variable, that evaluates to different expressions in different contexts, it *feels like* reality itself is unstable—what your mind's eye sees, shifts around depending on where it looks.

This often confuses undergraduates (and postmodernist professors) who discover a sentence with more than one interpretation; they think they have discovered an unstable portion of reality.

“Oh my gosh! ‘The Sun goes around the Earth’ is true for Hunga Huntergatherer, but for Amara Astronomer, ‘The Sun goes around the Earth’ is false! There is no fixed truth!” The deconstruction of this sophomoric nitwittery is left as an exercise to the reader.

And yet, even I initially found myself writing “If X is 5 on some occasions and 4 on another, the sentence ‘ $2 + 2 = X$ ’ may have no fixed truth-value.” There is not *one* sentence with a *variable* truth-value. “ $2 + 2 = X$ ” *has no* truth-value. It is not a *proposition*, not yet, not as mathematicians define proposition-ness, any more than “ $2 + 2 =$ ” is a proposition, or “Fred jumped over the” is a grammatical sentence.

But this fallacy tends to sneak in, even when you allegedly know better, because, well, that's how the algorithm feels from inside<sup>7</sup>.

---

6. Page 223, ‘How An Algorithm Feels From Inside’.

7. Page 223, ‘How An Algorithm Feels From Inside’.

## 28. 37 Ways That Words Can Be Wrong<sup>1</sup>

**Followup to:** Just about every post in February, and some in March

Some reader is bound to declare that a better title for this post would be “37 Ways That You Can Use Words Unwisely”, or “37 Ways That Suboptimal Use Of Categories Can Have Negative Side Effects On Your Cognition”.

But one of the primary lessons of this gigantic list is that saying “There’s no way my choice of X can be ‘wrong’” is nearly always an error in practice, whatever the theory. You can always be wrong. Even when it’s theoretically impossible to be wrong, you can still be wrong. There is never a Get-Out-Of-Jail-Free card for anything you do. That’s life.

Besides, I can define the word “wrong” to mean anything I like - it’s not like a word can be *wrong*.

Personally, I think it quite justified to use the word “wrong” when:

1. *A word fails to connect to reality in the first place.* Is Socrates a framster? Yes or no? (The Parable of the Dagger.<sup>2</sup>)
2. *Your argument, if it worked, could coerce reality to go a different way by choosing a different word definition.* Socrates is a human, and humans, by definition, are mortal. So if you defined humans to not be mortal, would Socrates live forever? (The Parable of Hemlock<sup>3</sup>.)
3. *You try to establish any sort of empirical proposition as being true “by definition”.* Socrates is a human, and humans, by definition, are mortal. So is it a *logical truth* if we empirically predict that Socrates should keel over if he drinks hemlock? It seems like there are logically possible, non-self-contradictory worlds where Socrates doesn’t keel over - where he’s immune to hemlock by a quirk of biochemistry, say. Logical truths are true in all possible worlds, and so never tell you *which* possible world you live

---

1. [http://lesswrong.com/lw/od/37\\_ways\\_that\\_words\\_can\\_be\\_wrong/](http://lesswrong.com/lw/od/37_ways_that_words_can_be_wrong/)

2. Page 191, ‘The Parable of the Dagger’.

3. Page 193, ‘The Parable of Hemlock’.

- in - and anything you can establish “by definition” is a logical truth. (The Parable of Hemlock<sup>4</sup>.)
4. *You unconsciously slap the conventional label on something, without actually using the verbal definition you just gave.* You know perfectly well that Bob is “human”, even though, on your definition, you can never call Bob “human” without first observing him to be mortal. (The Parable of Hemlock<sup>5</sup>.)
  5. *The act of labeling something with a word, disguises a challengable inductive inference you are making.* If the last 11 egg-shaped objects drawn have been blue, and the last 8 cubes drawn have been red, it is a matter of induction to say this rule will hold in the future. But if you call the blue eggs “bleggs” and the red cubes “rubes”, you may reach into the barrel, feel an egg shape, and think “Oh, a blegg.” (Words as Hidden Inferences.<sup>6</sup>)
  6. *You try to define a word using words, in turn defined with ever-more-abstract words, without being able to point to an example.* “What is red?” “Red is a color.” “What’s a color?” “It’s a property of a thing?” “What’s a thing? What’s a property?” It never occurs to you to point to a stop sign and an apple. (Extensions and Intensions.<sup>7</sup>)
  7. *The extension doesn’t match the intension.* We aren’t consciously aware of our identification of a red light in the sky as “Mars”, which will probably happen regardless of your attempt to define “Mars” as “The God of War”. (Extensions and Intensions.<sup>8</sup>)
  8. *Your verbal definition doesn’t capture more than a tiny fraction of the category’s shared characteristics, but you try to reason as if it does.* When the philosophers of Plato’s Academy claimed that the best definition of a human was a “featherless biped”, Diogenes the Cynic is said to have exhibited a plucked chicken and declared “Here is Plato’s Man.” The Platonists promptly changed their definition

---

4. Page 193, ‘The Parable of Hemlock’.

5. Page 193, ‘The Parable of Hemlock’.

6. Page 197, ‘Words as Hidden Inferences’.

7. Page 200, ‘Extensions and Intensions’.

8. Page 200, ‘Extensions and Intensions’.

- to “a featherless biped with broad nails”. (Similarity Clusters.<sup>9</sup>)
9. *You try to treat category membership as all-or-nothing, ignoring the existence of more and less typical subclusters.* Ducks and penguins are less typical birds than robins and pigeons. Interestingly, a between-groups experiment showed that subjects thought a disease was more likely to spread from robins to ducks on an island, than from ducks to robins. (Typicality and Asymmetrical Similarity.<sup>10</sup>)
  10. *A verbal definition works well enough in practice to point out the intended cluster of similar things, but you nitpick exceptions.* Not every human has ten fingers, or wears clothes, or uses language; but if you look for an empirical cluster of things which share these characteristics, you’ll get enough information that the occasional nine-fingered human won’t fool you. (The Cluster Structure of Thingspace<sup>11</sup>.)
  11. *You ask whether something “is” or “is not” a category member but can’t name the question you really want answered.* What is a “man”? Is Barney the Baby Boy a “man”? The “correct” answer may depend considerably on whether the query you *really* want answered is “Would hemlock be a good thing to feed Barney?” or “Will Barney make a good husband?” (Disguised Queries.<sup>12</sup>)
  12. *You treat intuitively perceived hierarchical categories like the only correct way to parse the world, without realizing that other forms of statistical inference are possible even though your brain doesn’t use them.* It’s much easier *for a human* to notice whether an object is a “blegg” or “rube”; than *for a human* to notice that red objects never glow in the dark, but red furred objects have all the other characteristics of bleggs. Other statistical algorithms work differently. (Neural Categories.<sup>13</sup>)
  13. *You talk about categories as if they are manna fallen from the Platonic Realm, rather than inferences implemented in a real brain.* The ancient philosophers said “Socrates is a man”,

---

9. Page 204, ‘Similarity Clusters’.

10. Page 206, ‘Typicality and Asymmetrical Similarity’.

11. Page 209, ‘The Cluster Structure of Thingspace’.

12. Page 213, ‘Disguised Queries’.

13. Page 217, ‘Neural Categories’.

- not, “My brain perceptually classifies Socrates as a match against the ‘human’ concept”. (How An Algorithm Feels From Inside.<sup>14</sup>)
14. *You argue about a category membership even after screening off all questions that could possibly depend on a category-based inference.* After you observe that an object is blue, egg-shaped, furred, flexible, opaque, luminescent, and palladium-containing, what’s left to ask by arguing, “Is it a blegg?” But if your brain’s categorizing neural network contains a (metaphorical) central unit corresponding to the inference of blegg-ness, it may still feel like there’s a leftover question. (How An Algorithm Feels From Inside.<sup>15</sup>)
  15. *You allow an argument to slide into being about definitions, even though it isn’t what you originally wanted to argue about.* If, before a dispute started about whether a tree falling in a deserted forest makes a “sound”, you asked the two soon-to-be arguers whether they thought a “sound” should be defined as “acoustic vibrations” or “auditory experiences”, they’d probably tell you to flip a coin. Only after the argument starts does the definition of a word become politically charged. (Disputing Definitions.<sup>16</sup>)
  16. *You think a word has a meaning, as a property of the word itself; rather than there being a label that your brain associates to a particular concept.* When someone shouts, “Yikes! A tiger!”, evolution would not favor an organism that thinks, “Hm... I have just heard the syllables ‘Tie’ and ‘Grr’ which my fellow tribemembers associate with their internal analogues of my own *tiger* concept and which *aiiiieee* CRUNCH CRUNCH GULP.” So the brain takes a shortcut, and it seems that the meaning of tigerness is a property of the label itself. People argue about the *correct meaning* of a label like “sound”. (Feel the Meaning.<sup>17</sup>)
  17. *You argue over the meanings of a word, even after all sides understand perfectly well what the other sides are trying to say.* The human ability to associate labels to concepts is a tool

---

14. Page 223, ‘How An Algorithm Feels From Inside’.

15. Page 223, ‘How An Algorithm Feels From Inside’.

16. Page 228, ‘Disputing Definitions’.

17. Page 234, ‘Feel the Meaning’.

for communication. When people *want* to communicate, we're hard to stop; if we have no common language, we'll draw pictures in sand. When you each understand what is in the other's mind, you are *done*. (The Argument From Common Usage.<sup>18</sup>)

18. *You pull out a dictionary in the middle of an empirical or moral argument.* Dictionary editors are historians of usage, not legislators of language. If the common definition contains a problem - if "Mars" is defined as the God of War, or a "dolphin" is defined as a kind of fish, or "Negroes" are defined as a separate category from humans, the dictionary will reflect the standard mistake. (The Argument From Common Usage.<sup>19</sup>)
19. *You pull out a dictionary in the middle of any argument ever.* Seriously, what the heck makes you think that dictionary editors are an authority on whether "atheism" is a "religion" or whatever? If you have any substantive issue whatsoever at stake, do you really think dictionary editors have access to ultimate wisdom that settles the argument? (The Argument From Common Usage.<sup>20</sup>)
20. *You defy common usage without a reason, making it gratuitously hard for others to understand you.* Fast stand up plutonium, with bagels without handle. (The Argument From Common Usage.<sup>21</sup>)
21. *You use complex renamings to create the illusion of inference.* Is a "human" defined as a "mortal featherless biped"? Then write: "All [mortal featherless bipeds] are mortal; Socrates is a [mortal featherless biped]; therefore, Socrates is mortal." Looks less impressive that way, doesn't it? (Empty Labels.<sup>22</sup>)
22. *You get into arguments that you could avoid if you just didn't use the word.* If Albert and Barry aren't allowed to use the word "sound", then Albert will have to say "A tree falling in a deserted forest generates acoustic vibrations", and Barry will say "A tree falling in a deserted forest generates

---

18. Page 238, 'The Argument from Common Usage'.

19. Page 238, 'The Argument from Common Usage'.

20. Page 238, 'The Argument from Common Usage'.

21. Page 238, 'The Argument from Common Usage'.

22. Page 242, 'Empty Labels'.

- no auditory experiences". When a word poses a problem, the simplest solution is to eliminate the word and its synonyms. (*Taboo Your Words*.<sup>23</sup>)
- 23. *The existence of a neat little word prevents you from seeing the details of the thing you're trying to think about.* What actually goes on in schools once you stop calling it "education"? What's a degree, once you stop calling it a "degree"? If a coin lands "heads", what's its radial orientation? What is "truth", if you can't say "accurate" or "correct" or "represent" or "reflect" or "semantic" or "believe" or "knowledge" or "map" or "real" or any other simple term? (Replace the Symbol with the Substance.<sup>24</sup>)
  - 24. *You have only one word, but there are two or more different things-in-reality, so that all the facts about them get dumped into a single undifferentiated mental bucket.* It's part of a detective's ordinary work to observe that Carol wore red last night, or that she has black hair; and it's part of a detective's ordinary work to wonder if maybe Carol dyes her hair. But it takes a subtler detective to wonder if there are two Carols, so that the Carol who wore red is not the same as the Carol who had black hair. (Fallacies of Compression.<sup>25</sup>)
  - 25. *You see patterns where none exist, harvesting other characteristics from your definitions even when there is no similarity along that dimension.* In Japan, it is thought that people of blood type A are earnest and creative, blood type Bs are wild and cheerful, blood type Os are agreeable and sociable, and blood type ABs are cool and controlled. (Categorizing Has Consequences.<sup>26</sup>)
  - 26. *You try to sneak in the connotations of a word, by arguing from a definition that doesn't include the connotations.* A "wiggin" is defined in the dictionary as a person with green eyes and black hair. The word "wiggin" also carries the connotation of someone who commits crimes and launches cute baby squirrels, but that part isn't in the dictionary. So you point to someone and say: "Green

---

23. Page 245, 'Taboo Your Words'.

24. Page 248, 'Replace the Symbol with the Substance'.

25. Page 253, 'Fallacies of Compression'.

26. Page 257, 'Categorizing Has Consequences'.

eyes? Black hair? See, told you he's a wigglin! Watch, next he's going to steal the silverware." (Sneaking in Connotations.<sup>27</sup>)

27. *You claim "X, by definition, is a Y!" On such occasions you're almost certainly trying to sneak in a connotation of Y that wasn't in your given definition.* You define "human" as a "featherless biped", and point to Socrates and say, "No feathers - two legs - he must be human!" But what you *really* care about is something else, like mortality. If what was in dispute was Socrates's number of legs, the other fellow would just reply, "Whaddaya mean, Socrates's got two legs? That's what we're arguing about in the first place!" (Arguing "By Definition".<sup>28</sup>)
28. *You claim "Ps, by definition, are Qs!"* If you see Socrates out in the field with some biologists, gathering herbs that might confer resistance to hemlock, there's no point in arguing "Men, by definition, are mortal!" The main time you feel the need to tighten the vise by insisting that something is true "by definition" is when there's other information that calls the default inference into doubt. (Arguing "By Definition".<sup>29</sup>)
29. *You try to establish membership in an empirical cluster "by definition".* You wouldn't feel the need to say, "Hinduism, by definition, is a religion!" because, well, of course Hinduism is a religion. It's not just a religion "by definition", it's, like, an *actual* religion. Atheism does not resemble the central members of the "religion" cluster, so if it wasn't for the fact that atheism is a religion *by definition*, you might go around thinking that atheism *wasn't* a religion. That's why you've got to crush all opposition by pointing out that "Atheism is a religion" is true *by definition*, because it isn't true any other way. (Arguing "By Definition".<sup>30</sup>)
30. *Your definition draws a boundary around things that don't really belong together.* You can claim, if you like, that you are *defining* the word "fish" to refer to salmon, guppies, sharks,

---

27. Page 260, 'Sneaking in Connotations'.

28. Page 263, 'Arguing "By Definition"'.

29. Page 263, 'Arguing "By Definition"'.

30. Page 263, 'Arguing "By Definition"'.

dolphins, and trout, but not jellyfish or algae. You can claim, if you like, that this is merely a list, and there is no way a list can be “wrong”. Or you can stop playing nitwit games and admit that you made a mistake and that dolphins don’t belong on the fish list. (Where to Draw the Boundary?<sup>31</sup>)

31. *You use a short word for something that you won’t need to describe often, or a long word for something you’ll need to describe often. This can result in inefficient thinking, or even misapplications of Occam’s Razor, if your mind thinks that short sentences sound “simpler”.* Which sounds more plausible, “God did a miracle” or “A supernatural universe-creating entity temporarily suspended the laws of physics”? (Entropy, and Short Codes.<sup>32</sup>)
32. *You draw your boundary around a volume of space where there is no greater-than-usual density, meaning that the associated word does not correspond to any performable Bayesian inferences.* Since green-eyed people are not more likely to have black hair, or vice versa, and they don’t share any other characteristics in common, why have a word for “wiggin”? (Mutual Information, and Density in Thingspace.<sup>33</sup>)
33. *You draw an unsimple boundary without any reason to do so.* The act of defining a word to refer to all humans, except black people, seems kind of suspicious. If you don’t present reasons to draw that particular boundary, trying to create an “arbitrary” word in that location is like a detective saying: “Well, I haven’t the slightest shred of support one way or the other for who could’ve murdered those orphans... but have we considered John Q. Wiffleheim as a suspect?” (Superexponential Conceptspace, and Simple Words.<sup>34</sup>)
34. *You use categorization to make inferences about properties that don’t have the appropriate empirical structure, namely, conditional independence given knowledge of the class, to be well-approximated by Naive Bayes.* No way am I trying to

---

<sup>31</sup>. Page 267, ‘Where to Draw the Boundary?’.

<sup>32</sup>. Page 271, ‘Entropy, and Short Codes’.

<sup>33</sup>. Page 275, ‘Mutual Information, and Density in Thingspace’.

<sup>34</sup>. Page 282, ‘Superexponential Conceptspace, and Simple Words’.

summarize this one. Just read the blog post. (Conditional Independence, and Naive Bayes.<sup>35</sup>)

- 35. *You think that words are like tiny little LISP symbols in your mind, rather than words being labels that act as handles to direct complex mental paintbrushes that can paint detailed pictures in your sensory workspace.* Visualize a “triangular lightbulb”. What did you see? (Words as Mental Paintbrush Handles.<sup>36</sup>)
- 36. *You use a word that has different meanings in different places as though it meant the same thing on each occasion, possibly creating the illusion of something protean and shifting.* “Martin told Bob the building was on his left.” But “left” is a function-word that evaluates with a speaker-dependent variable grabbed from the surrounding context. Whose “left” is meant, Bob’s or Martin’s? (Variable Question Fallacies.<sup>37</sup>)
- 37. *You think that definitions can’t be “wrong”, or that “I can define a word any way I like!”* This kind of attitude teaches you to indignantly defend your past actions, instead of paying attention to their consequences, or fessing up to your mistakes. (37 Ways That Suboptimal Use Of Categories Can Have Negative Side Effects On Your Cognition.<sup>38</sup>)

Everything you do in the mind has an effect, and your brain races ahead unconsciously without your supervision.

Saying “Words are arbitrary; I can define a word any way I like” makes around as much sense as driving a car over thin ice with the accelerator floored and saying, “Looking at this steering wheel, I can’t see why one radial angle is special - so I can turn the steering wheel any way I like.”

If you’re trying to go anywhere, or even just trying to *survive*, you had better start paying attention to the three or six dozen optimality criteria that control how you use words, definitions, categories, classes, boundaries, labels, and concepts.

---

<sup>35</sup>. Page 288, ‘Conditional Independence, and Naive Bayes’.

<sup>36</sup>. Page 295, ‘Words as Mental Paintbrush Handles’.

<sup>37</sup>. Page 298, ‘Variable Question Fallacies’.

<sup>38</sup>. Page 301, ‘37 Ways That Words Can Be Wrong’.



## **Part IV**

# **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.*



## I. A Fable of Science and Politics<sup>1</sup>

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 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

---

<sup>1</sup>. [http://lesswrong.com/lw/gt/a\\_fable\\_of\\_science\\_and\\_politics/](http://lesswrong.com/lw/gt/a_fable_of_science_and_politics/)

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 tremor 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<sup>1</sup>

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<sup>2</sup> position on an issue.

In Artificial Intelligence, and particularly in the domain of non-monotonic 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?”

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 enter-

---

1. [http://lesswrong.com/lw/gw/politics\\_is\\_the\\_mindkiller/](http://lesswrong.com/lw/gw/politics_is_the_mindkiller/)

2. Page 315, ‘A Fable of Science and Politics’.

ing 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<sup>3</sup>. 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<sup>4</sup>. 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 or politics<sup>5</sup>.

(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 315, 'A Fable of Science and Politics'.

4. [http://en.wikipedia.org/wiki/Wikipedia:Neutral\\_point\\_of\\_view](http://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view)

5. Page 315, 'A Fable of Science and Politics'.

### **3. Policy Debates Should Not Appear One-Sided<sup>1</sup>**

Robin Hanson recently proposed stores where banned products could be sold<sup>2</sup>. 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.<sup>3</sup> 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.

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

---

1. [http://lesswrong.com/lw/gz/policy\\_debates\\_should\\_not\\_appear\\_onesided/](http://lesswrong.com/lw/gz/policy_debates_should_not_appear_onesided/)

2. [http://www.overcomingbias.com/2007/03/paternalism\\_is\\_.html](http://www.overcomingbias.com/2007/03/paternalism_is_.html)

3. Page 319, ‘Politics is the Mind-Killer’.

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 polysyllabic 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<sup>4</sup> and the just-world fallacy<sup>5</sup>.

4. Page 363, ‘The Affect Heuristic’.

5. [http://en.wikipedia.org/wiki/Just-world\\_fallacy](http://en.wikipedia.org/wiki/Just-world_fallacy)

## 4. The Scales of Justice, the Notebook of Rationality<sup>1</sup>

Lady Justice<sup>2</sup> 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<sup>3</sup>, 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"<sup>4</sup> 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 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 phys-

---

1. [http://lesswrong.com/lw/b1/the\\_scales\\_of\\_justice\\_the\\_notebook\\_of\\_rationality/](http://lesswrong.com/lw/b1/the_scales_of_justice_the_notebook_of_rationality/)

2. [http://en.wikipedia.org/wiki/Lady\\_Justice](http://en.wikipedia.org/wiki/Lady_Justice)

3. Page 315, 'A Fable of Science and Politics'.

4. [http://www-abc.mpib-berlin.mpg.de/users/r20/finucane00\\_the\\_affect\\_heuristic.pdf](http://www-abc.mpib-berlin.mpg.de/users/r20/finucane00_the_affect_heuristic.pdf)

ical 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<sup>1</sup>

*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<sup>2</sup>

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 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.)

---

1. [http://lesswrong.com/lw/bz/correspondence\\_bias/](http://lesswrong.com/lw/bz/correspondence_bias/)

2. <http://www.wjh.harvard.edu/~dtg/>

Gilbert%20&%20Malone%20(CORRESPONDENCE%20BIAS).pdf

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<sup>3</sup>.

And yet people sometimes ask me why I want to save the world<sup>4</sup>. 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.

---

3. <http://www.singinst.org/blog/2007/06/16/transhumanism-as-simplified-humanism/>

4. <http://singinst.org/AIRisk.pdf>

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<sup>5</sup> by human events.

---

5. [http://lesswrong.com/lw/bs/think\\_like\\_reality/](http://lesswrong.com/lw/bs/think_like_reality/)

## 6. Are Your Enemies Innately Evil?<sup>1</sup>

### Followup to: Correspondence Bias<sup>2</sup>

As previously discussed<sup>3</sup>, 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 bizarreness<sup>4</sup> 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.

But politics is the mind-killer.<sup>5</sup> 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.

---

1. [http://lesswrong.com/lw/io/are\\_your\\_enemies\\_innately\\_evil/](http://lesswrong.com/lw/io/are_your_enemies_innately_evil/)

2. Page 326, 'Correspondence Bias'.

3. Page 326, 'Correspondence Bias'.

4. [http://lesswrong.com/lw/hs/think\\_like\\_reality/](http://lesswrong.com/lw/hs/think_like_reality/)

5. Page 319, 'Politics is the Mind-Killer'.

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 battlefield. 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<sup>6</sup> 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 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

---

6. Page 321, ‘Policy Debates Should Not Appear One-Sided’.

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<sup>7</sup>. 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<sup>8</sup>).

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 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. Page 319, 'Politics is the Mind-Killer'.

8. [http://www.youtube.com/watch?v=y\\_EKHKrCzIE](http://www.youtube.com/watch?v=y_EKHKrCzIE)

## 7. The Robbers Cave Experiment<sup>1</sup>

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<sup>2</sup> 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 began hurling insults. They named themselves the Rattlers and the Eagles (they hadn’t needed names when they were the only group on the camp-ground).

When the contests and prizes were announced, in accordance with pre-established experimental procedure, the intergroup rivalry

---

1. [http://lesswrong.com/lw/lt/the\\_robbers\\_cave\\_experiment/](http://lesswrong.com/lw/lt/the_robbers_cave_experiment/)

2. <http://psychclassics.yorku.ca/Sherif/>

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<sup>3</sup>, 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*.<sup>4</sup> University of Oklahoma.

3. [http://lesswrong.com/lw/lx/evolutionary\\_psychology/](http://lesswrong.com/lw/lx/evolutionary_psychology/)

4. <http://psychclassics.yorku.ca/Sherif/>

## 8. Reversed Stupidity Is Not Intelligence<sup>1</sup>

“...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<sup>2</sup>, 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.  $p(\text{cults}|\text{aliens})$  isn’t less than  $p(\text{cults}|\neg\text{aliens})$ , unless you suppose that poorly hidden aliens would deliberately suppress flying saucer cults. By the Bayesian definition of evidence<sup>3</sup>, the observation “flying 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 re-

---

1. [http://lesswrong.com/lw/lw/reversed\\_stupidity\\_is\\_not\\_intelligence/](http://lesswrong.com/lw/lw/reversed_stupidity_is_not_intelligence/)

2. Page 406, ‘Every Cause Wants To Be A Cult’.

3. Page 28, ‘An Intuitive Explanation of Bayes’ Theorem’.

versing 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<sup>4</sup> 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<sup>5</sup> is the only valid one.
- 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<sup>6</sup>.
- 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<sup>7</sup> argument is not valid.

---

4. Page 374, 'The Halo Effect'.

5. [http://lesswrong.com/lw/2k/the\\_least\\_convenient\\_possible\\_world/](http://lesswrong.com/lw/2k/the_least_convenient_possible_world/)

6. [http://lesswrong.com/lw/r5/the\\_quantum\\_physics\\_sequence/](http://lesswrong.com/lw/r5/the_quantum_physics_sequence/)

7. <http://plover.net/~bonds/adhominem.html>

- 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.
- If a hundred inventors fail<sup>8</sup> 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<sup>9</sup>.

---

8. [http://lesswrong.com/lw/vs/selling\\_nonapples/](http://lesswrong.com/lw/vs/selling_nonapples/)

9. [http://lesswrong.com/lw/l9/artificial\\_addition/](http://lesswrong.com/lw/l9/artificial_addition/)

## 9. Argument Screens Off Authority<sup>1</sup>

Black Belt Bayesian<sup>2</sup> (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<sup>3</sup>, 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 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 noise—it seems that Arthur should view David as having a great ad-

---

1. [http://lesswrong.com/lw/lx/argument\\_screens\\_off\\_authority/](http://lesswrong.com/lw/lx/argument_screens_off_authority/)

2. <http://www.acceleratingfuture.com/steven/?p=33>

3. Page 335, ‘Reversed Stupidity Is Not Intelligence’.

vantage 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<sup>4</sup>, a difference unaccountable in the boringly clean methods of Bayesian probability theory<sup>5</sup>. 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?

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  $p(H|E_1) = 90\%$  and  $p(H|E_2) = 9\%$ , what is the probability  $p(H|E_1, E_2)$ ? If learning  $E_1$  is true leads us to assign 90% probability to  $H$ , and learning  $E_2$  is true leads us to assign 9% probability to  $H$ , then what probability should we assign to  $H$  if we learn both  $E_1$  and  $E_2$ ? This is simply not something you can calculate in probability theory from the information given. No, the missing information is

4. Page 86, ‘What is Evidence?’.

5. Page 28, ‘An Intuitive Explanation of Bayes' Theorem’.

not the prior probability of H. E<sub>1</sub> and E<sub>2</sub> may not be independent of each other.

Suppose that H is “My sidewalk is slippery”, E<sub>1</sub> is “My sprinkler is running”, and E<sub>2</sub> 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<sup>6</sup> and Causality<sup>7</sup>. (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:

---

6. <http://www.amazon.com/Probabilistic-Reasoning-Intelligent-Systems-Plausible/dp/1558604790/>

7. <http://www.amazon.com/Causality-Reasoning-Inference-Judea-Pearl/dp/0521773628/>

$$\begin{aligned} p(\text{die1}, \text{die2}) &= p(\text{die1}) * p(\text{die2}) \\ p(\text{die1}, \text{die2}|\text{sum}) &\neq p(\text{die1}|\text{sum}) * p(\text{die2}|\text{sum}) \end{aligned}$$

If you look at the correct sidewalk diagram, you see facts like:

$$\begin{aligned} p(\text{slippery}|\text{night}) &\neq p(\text{slippery}) \\ p(\text{slippery}|\text{sprinkler}) &\neq p(\text{slippery}) \\ p(\text{slippery}|\text{night, sprinkler}) &= p(\text{slippery}|\text{sprinkler}) \end{aligned}$$

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  $\rightarrow$  Argument Goodness  $\rightarrow$  Expert Belief

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:

$$p(\text{truth}|\text{argument, expert}) = p(\text{truth}|\text{argument})$$

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 dif-

ferent 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<sup>1</sup>

### Continuation of: Argument Screens Off Authority<sup>2</sup>

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<sup>3</sup> 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 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<sup>4</sup>, because it will distract you from more direct arguments. It’s

---

1. [http://lesswrong.com/lw/lj/bug\\_the\\_query/](http://lesswrong.com/lw/lj/bug_the_query/)

2. Page 338, ‘Argument Screens Off Authority’.

3. Page 338, ‘Argument Screens Off Authority’.

4. Page 505, ‘Knowing About Biases Can Hurt People’.

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.<sup>5</sup>

Just as you can't always experiment today<sup>6</sup>, 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!*

---

5. Page 335, 'Reversed Stupidity Is Not Intelligence'.

6. [http://lesswrong.com/lw/io/is\\_molecular\\_nanotechnology\\_scientific/](http://lesswrong.com/lw/io/is_molecular_nanotechnology_scientific/)

## 11. Rationality and the English Language<sup>1</sup>

Yesterday<sup>2</sup>, someone said that my writing reminded them of George Orwell's Politics and the English Language<sup>3</sup>. I was honored. Especially since I'd already thought of today's topic.

If you *really* want an artist's perspective<sup>4</sup> 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*.

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

---

1. [http://lesswrong.com/lw/jc/rationality\\_and\\_the\\_english\\_language/](http://lesswrong.com/lw/jc/rationality_and_the_english_language/)

2. [http://lesswrong.com/lw/jb/applause\\_lights/fit](http://lesswrong.com/lw/jb/applause_lights/fit)

3. <http://www.k-t.com/Orwell/index.cgi/work/essays/language.html>

4. [http://lesswrong.com/lw/ja/we\\_dont\\_really\\_want\\_your\\_participation/](http://lesswrong.com/lw/ja/we_dont_really_want_your_participation/)

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<sup>5</sup>, 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, rephrasing 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 authoritarianism, 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.

---

5. Page 180, ‘Applause Lights’.

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<sup>6</sup> 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<sup>7</sup> 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 clichés, 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 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

---

6. Page 151, 'Semantic Stopsigns'.

7. Page 180, 'Applause Lights'.

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<sup>8</sup> 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=vDfjyvFZotS&sig=sZtZG8\\_yfTFIUpCIq\\_YSUgR749M#P](http://books.google.com/books?id=NK2dLn48zWIC&pg=PA338&lpg=PA338&ots=vDfjyvFZotS&sig=sZtZG8_yfTFIUpCIq_YSUgR749M#P)

## 12. The Litany Against Gurus<sup>1</sup>

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.

---

<sup>1</sup>. [http://lesswrong.com/lw/m2/the\\_litany\\_against\\_gurus/](http://lesswrong.com/lw/m2/the_litany_against_gurus/)

## 13. Politics and Awful Art<sup>1</sup>

### Followup to: Rationality and the English Language<sup>2</sup>

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<sup>3</sup>, 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

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"

1. [http://lesswrong.com/lw/m3/politics\\_and\\_awful\\_art/](http://lesswrong.com/lw/m3/politics_and_awful_art/)

2. Page 345, 'Rationality and the English Language'.

3. Page 349, 'The Litany Against Gurus'.

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<sup>4</sup> entering the supercritical phase<sup>5</sup> 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<sup>6</sup> philosophical<sup>7</sup> lectures<sup>8</sup> into the climax of your fictional novel...

---

4. Page 384, ‘Affective Death Spirals’.

5. Page 392, ‘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](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](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](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<sup>1</sup>

### Followup to: Politics and Awful Art<sup>2</sup>

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<sup>3</sup>.

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<sup>4</sup>. 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.

(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 enormous—

---

1. [http://lesswrong.com/lw/m5/false\\_laughter/](http://lesswrong.com/lw/m5/false_laughter/)

2. Page 350, 'Politics and Awful Art'.

3. Page 350, 'Politics and Awful Art'.

4. Page 451, 'Cached Thoughts'.

ly 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 funny 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<sup>5</sup> 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 non-cultishness—well, like most kinds of obligatory laughter, it doesn't quite

---

5. Page 415, 'Guardians of Ayn Rand'.

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.

## 15. Human Evil and Muddled Thinking<sup>1</sup>

**Followup to:** Rationality and the English Language<sup>2</sup>

George Orwell<sup>3</sup> 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<sup>4</sup> 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<sup>5</sup>; it is hard for you to take seriously<sup>6</sup>. 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 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*...

---

1. [http://lesswrong.com/lw/jd/human\\_evil\\_and\\_muddled\\_thinking/](http://lesswrong.com/lw/jd/human_evil_and_muddled_thinking/)

2. Page 345, 'Rationality and the English Language'.

3. <http://www.k-1.com/Orwell/index.cgi/work/essays/language.html>

4. Page 172, 'Making History Available'.

5. Page 172, 'Making History Available'.

6. Page 172, 'Making History Available'.

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<sup>7</sup> the planned famines of Stalin and Mao, which killed tens of millions<sup>8</sup>. 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 air-brushing 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:

'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?'

7. Page 172, 'Making History Available'.

8. [http://lesswrong.com/lw/bw/scope\\_insensitivity/](http://lesswrong.com/lw/bw/scope_insensitivity/)

‘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<sup>9</sup>—who don’t think that overcoming bias is important. This is your *mind* we’re talking about. Your human intelligence<sup>10</sup>. It separates you from an ape. It built this world. You don’t think how the mind works 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<sup>11</sup> is on the same level with trying to create plans that will *really* save as many lives as possible?

Orwell<sup>12</sup> 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

---

9. <http://www.marginalrevolution.com/marginalrevolution/2007/08/how-important-i.html>

10. <http://www.singinst.org/blog/2007/07/10/the-power-of-intelligence/>

11. [http://lesswrong.com/lw/hw/scope\\_insensitivity/](http://lesswrong.com/lw/hw/scope_insensitivity/)

12. <http://www.k-t.com/Orwell/index.cgi/work/essays/language.html>

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. 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.*



## I. The Affect Heuristic<sup>1</sup>

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.

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).

---

<sup>1</sup>. [http://lesswrong.com/lw/lg/the\\_affect\\_heuristic/](http://lesswrong.com/lw/lg/the_affect_heuristic/)

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.”<sup>2</sup>

---

Denes-Raj, V., & Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their 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.<sup>3</sup> *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.<sup>4</sup> *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

---

2. <http://heuristics.behaviouralfinance.net/affect/Slovic.pdf>

3. [http://www-abc.mpib-berlin.mpg.de/users/r20/finucane00\\_the\\_affect\\_heuristic.pdf](http://www-abc.mpib-berlin.mpg.de/users/r20/finucane00_the_affect_heuristic.pdf)

4. <http://faculty.chicagogs.edu/christopher.hsee/vita/Papers/AffectionEffectInInsurance.pdf>

Heuristic for Behavioral Economics.<sup>5</sup> *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.

---

5. <http://heuristics.behaviouralfinance.net/affect/Slovoz.pdf>

## 2. Evaluability (And Cheap Holiday Shopping)<sup>1</sup>

**Followup to:** The Affect Heuristic<sup>2</sup>

With the *expensive* part of the Hallowthankmas<sup>3</sup> 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.

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-evaluuable*.

---

1. [http://lesswrong.com/lw/lb/evaluability\\_and Cheap\\_holiday\\_shopping/](http://lesswrong.com/lw/lb/evaluability_and Cheap_holiday_shopping/)

2. Page 363, ‘The Affect Heuristic’.

3. <http://www.overcomingbias.com/2007/11/merry-hallowmas.html>

The torn cover, on the other hand—that stands out. That has a definite affective valence<sup>4</sup>: 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, than 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  $29/36$  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

---

4. Page 363, ‘The Affect Heuristic’.

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<sup>5</sup> 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?

Hsee1998

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

---

5. [http://lesswrong.com/lw/ks/the\\_wonder\\_of\\_evolution/](http://lesswrong.com/lw/ks/the_wonder_of_evolution/)

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.<sup>6</sup> *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.<sup>7</sup> *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.<sup>8</sup> *Journal of Socio-Economics*, 31: 329–342.

6. <http://faculty.chicagogs.edu/christopher.hsee/vita/Papers/EvaluabilityHypothesis.pdf>

7. <http://faculty.chicagogs.edu/christopher.hsee/vita/Papers/LessIsBetter.pdf>

8. <http://heuristics.behaviouralfinance.net/affect/Slov02.pdf>

### 3. Unbounded Scales, Huge Jury Awards, & Futurism<sup>1</sup>

**Followup to:** Evaluability<sup>2</sup>

“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?  $\circ$  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 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.

---

1. [http://lesswrong.com/lw/li/unbounded\\_scales\\_huge\\_jury\\_awards\\_futurism/](http://lesswrong.com/lw/li/unbounded_scales_huge_jury_awards_futurism/)

2. Page 367, ‘Evaluability (And Cheap Holiday Shopping)’.

“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<sup>3</sup>. *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.<sup>4</sup> *Journal of Risk and Uncertainty*, 19: 203-235.

3. [http://ist-socrates.berkeley.edu/~maccoun/LP\\_KahnemanSchkadeSunstein1998.pdf](http://ist-socrates.berkeley.edu/~maccoun/LP_KahnemanSchkadeSunstein1998.pdf)

4. <http://www.springerlink.com/content/u232267854514u6m/fulltext.pdf>

## 4. The Halo Effect<sup>1</sup>

The affect heuristic<sup>2</sup> 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<sup>3</sup> 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 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).

---

1. [http://lesswrong.com/lw/lj/the\\_halo\\_effect/](http://lesswrong.com/lw/lj/the_halo_effect/)

2. Page 363, ‘The Affect Heuristic’.

3. Page 363, ‘The Affect Heuristic’.

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 (Hamermesh & 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, Karabanic, & 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<sup>4</sup> or serene<sup>5</sup>. You should be suspicious if the people you know seem to separate too cleanly into devils and angels.

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., Karabanic, 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

---

4. <http://ozyandmillie.org/2003/03/24/ozy-and-millie-1134/>

5. <http://ozyandmillie.org/2006/11/16/ozy-and-millie-1770/>

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*, 35, 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.

Hamermesh, 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<sup>1</sup>

### Followup to: The Halo Effect<sup>2</sup>

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<sup>3</sup> 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:

“How tough can it be to act all brave and courageous  
when you're pretty much invulnerable?”

—*Empowered*, Vol. I

I can't remember if I read the following point somewhere, or hypothesized it myself: *Fame*, in particular, seems to combine addi-

---

1. [http://lesswrong.com/lw/lk/superhero\\_bias/](http://lesswrong.com/lw/lk/superhero_bias/)

2. Page 374, ‘The Halo Effect’.

3. Page 374, ‘The Halo Effect’.

tively 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<sup>4</sup>” may be a fine applause light<sup>5</sup>, 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

---

4. [http://lesswrong.com/lw/bx/one\\_life\\_against\\_the\\_world/](http://lesswrong.com/lw/bx/one_life_against_the_world/)

5. Page 180, ‘Applause Lights’.

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<sup>6</sup>.

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. [http://lesswrong.com/lw/le/lost\\_purposes/](http://lesswrong.com/lw/le/lost_purposes/)

## 6. Mere Messiahs<sup>1</sup>

### Followup to: Superhero Bias<sup>2</sup>

Yesterday I discussed how the halo effect<sup>3</sup>, 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<sup>4</sup>.)

“The police officer who puts their life on the line with no super-powers”, I said, “reveals far greater virtue than Superman, who is a *mere superhero*.”

But let’s be more specific.

John Perry<sup>5</sup> 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 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.

---

1. [http://lesswrong.com/lw/ll/mere\\_messiahs/](http://lesswrong.com/lw/ll/mere_messiahs/)

2. Page 378, ‘Superhero Bias’.

3. Page 374, ‘The Halo Effect’.

4. [http://lesswrong.com/lw/bx/one\\_life\\_against\\_the\\_world/](http://lesswrong.com/lw/bx/one_life_against_the_world/)

5. <http://www.nleomf.org/911heroes/Perry.html>

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, it only lasted a few hours—nothing compared to the duration, or even the intensity, of the tortures the Inquisition visited upon suspected witches. 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 a few hours' 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 be-

lieved to be right, confronted a church he believed to be corrupt, and died for it. Without benefit of hindsight<sup>6</sup>, 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.”

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<sup>7</sup> his soul would survive, he doesn’t deserve more honor than John Perry.

---

6. Page 133, ‘Hindsight bias’.

7. Page 108, ‘Belief in Belief’.

## 7. Affective Death Spirals<sup>1</sup>

**Followup to:** The Affect Heuristic<sup>2</sup>, The Halo Effect<sup>3</sup>

Many<sup>4</sup>, many<sup>5</sup>, many<sup>6</sup> 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.

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<sup>7</sup> is that any perceived positive characteristic (such as attractiveness or strength) increases perception of any oth-

---

1. [http://lesswrong.com/lw/lm/affective\\_death\\_spirals/](http://lesswrong.com/lw/lm/affective_death_spirals/)

2. Page 363, 'The Affect Heuristic'.

3. Page 374, 'The Halo Effect'.

4. Page 133, 'Hindsight bias'.

5. Page 147, 'Fake Causality'.

6. Page 519, 'Rationalization'.

7. Page 374, 'The Halo Effect'.

er positive characteristic (such as intelligence or courage). Even when it makes no sense, or less than no sense<sup>8</sup>.

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<sup>9</sup>.

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. Page 378, 'Superhero Bias'.

9. [http://www.youtube.com/watch?v=ORqcix3\\_Evg&feature=related](http://www.youtube.com/watch?v=ORqcix3_Evg&feature=related)

## 8. Resist the Happy Death Spiral<sup>1</sup>

### Followup to: Affective Death Spirals<sup>2</sup>

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<sup>3</sup> with respect to Science itself? The happy death spiral<sup>4</sup> starts when you believe something is *so* wonderful that the halo effect<sup>5</sup> 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...*

If you retrieve the standard<sup>6</sup> cached<sup>7</sup> deep wisdom<sup>8</sup> 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

---

1. [http://lesswrong.com/lw/ln/resist\\_the\\_happy\\_death\\_spiral/](http://lesswrong.com/lw/ln/resist_the_happy_death_spiral/)

2. Page 384, 'Affective Death Spirals'.

3. Page 384, 'Affective Death Spirals'.

4. Page 384, 'Affective Death Spirals'.

5. Page 374, 'The Halo Effect'.

6. Page 454, 'The "Outside the Box" Box'.

7. Page 451, 'Cached Thoughts'.

8. Page 465, 'How to Seem (and Be) Deep'.

or disprove<sup>9</sup> the dragon in my garage<sup>10</sup>.” 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<sup>11</sup>? Why would you pay attention to your own thoughts, if you knew you were trying to manipulate yourself<sup>12</sup>?

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.

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

9. [http://lesswrong.com/lw/i8/religions\\_claim\\_to\\_be\\_nondisprovable/](http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/)

10. Page 108, ‘Belief in Belief’.

11. Page 519, ‘Rationalization’.

12. Page 565, ‘Doublethink (Choosing to be Biased)’.

to fly. “Science has nothing to say about a parent’s love for their child”—sorry, that’s simply false<sup>13</sup>. 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<sup>14</sup>, 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.

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<sup>15</sup> 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<sup>16</sup>, which makes us more likely to accept future positive claims once we’ve accepted an initial

---

13. [http://lesswrong.com/lw/lv/evolutionary\\_psychology/](http://lesswrong.com/lw/lv/evolutionary_psychology/)

14. <http://www.singinst.org/blog/2007/10/21/should-ethicists-be-inside-or-outside-a-profession/>

15. Page 122, ‘The Virtue of Narrowness’.

16. Page 374, ‘The Halo Effect’.

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<sup>17</sup>, 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<sup>18</sup> or avoiding the real weak points<sup>19</sup>. Given the danger of the happy death spiral, it makes sense to try to avoid being happy about *unsettled* 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<sup>20</sup>—demanding sufficient<sup>21</sup> evidence to compensate for complexity<sup>22</sup>, searching<sup>23</sup> for flaws as well as support, invoking curiosity<sup>24</sup>—once they've accepted some core premise. Without the conjunction fallacy<sup>25</sup>, there might still be a halo effect<sup>26</sup>, but there wouldn't be a happy death spiral<sup>27</sup>.

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.

---

17. [http://lesswrong.com/lw/jk/burdensome\\_details/](http://lesswrong.com/lw/jk/burdensome_details/)

18. Page 511, 'One Argument Against An Army'.

19. Page 525, 'Avoiding Your Belief's Real Weak Points'.

20. [http://lesswrong.com/lw/jk/burdensome\\_details/](http://lesswrong.com/lw/jk/burdensome_details/)

21. Page 89, 'How Much Evidence Does It Take?'.

22. Page 95, 'Occam's Razor'.

23. Page 516, 'What Evidence Filtered Evidence?'.

24. Page 640, 'The Meditation on Curiosity'.

25. [http://lesswrong.com/lw/ji/conjunction\\_fallacy/](http://lesswrong.com/lw/ji/conjunction_fallacy/)

26. Page 374, 'The Halo Effect'.

27. Page 384, 'Affective Death Spirals'.

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.<sup>28</sup> Then the chain reaction goes *supercritical*. More on this tomorrow.

**Addendum:** Stuart Armstrong gives closely related advice:<sup>29</sup>

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)

---

28. Page 319, 'Politics is the Mind-Killer'.

29. [http://lesswrong.com/lw/lm/affective\\_death\\_spirals/gp5](http://lesswrong.com/lw/lm/affective_death_spirals/gp5)

conducting a biased search for negative points until you feel unhappy again (8) forcibly shoving an idea into a safe box.

## 9. Uncritical Supercriticality<sup>1</sup>

### Followup to: Resist the Happy Death Spiral<sup>2</sup>

Every now and then, you see people arguing over whether atheism is a “religion”. As I touched on in Purpose and Pragmatism<sup>3</sup>, 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<sup>4</sup>”, 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”.

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.

---

1. [http://lesswrong.com/lw/lo/uncritical\\_supercriticality/](http://lesswrong.com/lw/lo/uncritical_supercriticality/)

2. Page 386, ‘Resist the Happy Death Spiral’.

3. [http://lesswrong.com/lw/lf/purpose\\_and\\_pragmatism/](http://lesswrong.com/lw/lf/purpose_and_pragmatism/)

4. [http://en.wikipedia.org/wiki/No\\_true\\_Scotsman](http://en.wikipedia.org/wiki/No_true_Scotsman)

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<sup>5</sup> an affective death spiral<sup>6</sup> is the principle of “burdensome details<sup>7</sup>”—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<sup>8</sup>, 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.

---

5. Page 386, ‘Resist the Happy Death Spiral’.

6. Page 384, ‘Affective Death Spirals’.

7. [http://lesswrong.com/lw/jk/burdensome\\_details/](http://lesswrong.com/lw/jk/burdensome_details/)

8. Page 374, ‘The Halo Effect’.

The diametric opposite of this advice, which sends the halo effect *supercritical*, is when it feels wrong to argue against *any* positive claim about the Great Idea. Politics is the mind-killer<sup>9</sup>. 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<sup>10</sup>, 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;

...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<sup>11</sup> 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!<sup>12</sup>) 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<sup>13</sup> 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

9. Page 319, ‘Politics is the Mind-Killer’.

10. [http://lesswrong.com/lw/ks/the\\_wonder\\_of\\_evolution/](http://lesswrong.com/lw/ks/the_wonder_of_evolution/)

11. <http://richardcarrier.blogspot.com/2007/01/defining-supernatural.html>

12. [http://lesswrong.com/lw/lo/uncritical\\_supercriticality/ggb](http://lesswrong.com/lw/lo/uncritical_supercriticality/ggb)

13. Page 154, ‘Mysterious Answers to Mysterious Questions’.

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<sup>14</sup> that faith is positive, but lack Christianity's exclusionary scripture to 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<sup>15</sup> 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.

---

<sup>14</sup>. Page 451, 'Cached Thoughts'.

<sup>15</sup>. Page 89, 'How Much Evidence Does It Take?'.

## 10. Evaporative Cooling of Group Beliefs<sup>1</sup>

### Followup to: Uncritical Supercriticality<sup>2</sup>

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 space-fleet<sup>3</sup> on September 27, 1975. (The Wikipedia article<sup>4</sup> 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<sup>5</sup>, 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 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

1. [http://lesswrong.com/lw/lr/evaporative\\_cooling\\_of\\_group\\_beliefs/](http://lesswrong.com/lw/lr/evaporative_cooling_of_group_beliefs/)

2. Page 392, ‘Uncritical Supercriticality’.

3. [http://findarticles.com/p/articles/mi\\_moSOR/is\\_n2\\_v59/ai\\_20913876/pg\\_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](http://en.wikipedia.org/wiki/Unarius_Academy_of_Science)

5. [http://www.colorado.edu/physics/2000/bec/evap\\_cool.html](http://www.colorado.edu/physics/2000/bec/evap_cool.html)

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<sup>6</sup> 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 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

---

6. [http://www.colorado.edu/physics/2000/bec/evap\\_cool.html](http://www.colorado.edu/physics/2000/bec/evap_cool.html)

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<sup>7</sup> 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

---

<sup>7</sup>. Page 180, 'Applause Lights'.

out. This theory of Internet moderation may not have served me too well in practice, so take it with a grain of salt.

## 11. When None Dare Urge Restraint<sup>1</sup>

### Followup to: Uncritical Supercriticality<sup>2</sup>

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.

This is the even darker mirror of the happy death spiral<sup>3</sup>—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 state-

---

1. [http://lesswrong.com/lw/lz/when\\_none\\_dare\\_urge\\_restraint/](http://lesswrong.com/lw/lz/when_none_dare_urge_restraint/)

2. Page 392, ‘Uncritical Supercriticality’.

3. Page 384, ‘Affective Death Spirals’.

ments 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<sup>4</sup>“. 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 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<sup>5</sup>. The emotional shock

---

4. Page 329, ‘Are Your Enemies Innately Evil?’.

5. Page 180, ‘Applause Lights’.

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<sup>6</sup> redditors! You may also enjoy A Fable of Science and Politics<sup>7</sup> and Policy Debates Should Not Appear One-Sided<sup>8</sup>.

---

6. [http://lesswrong.com/lw/1/about\\_less\\_wrong/](http://lesswrong.com/lw/1/about_less_wrong/)

7. Page 315, ‘A Fable of Science and Politics’.

8. Page 321, ‘Policy Debates Should Not Appear One-Sided’.

## 12. The Robbers Cave Experiment<sup>1</sup>

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<sup>2</sup> 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 began hurling insults. They named themselves the Rattlers and the Eagles (they hadn’t needed names when they were the only group on the camp-ground).

When the contests and prizes were announced, in accordance with pre-established experimental procedure, the intergroup rivalry

---

1. [http://lesswrong.com/lw/lt/the\\_robbers\\_cave\\_experiment/](http://lesswrong.com/lw/lt/the_robbers_cave_experiment/)

2. <http://psychclassics.yorku.ca/Sherif/>

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<sup>3</sup>, 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*.<sup>4</sup> University of Oklahoma.

3. [http://lesswrong.com/lw/lx/evolutionary\\_psychology/](http://lesswrong.com/lw/lx/evolutionary_psychology/)

4. <http://psychclassics.yorku.ca/Sherif/>

## 13. Every Cause Wants To Be A Cult<sup>1</sup>

**Followup to:** Correspondence Bias<sup>2</sup>, Affective Death Spirals<sup>3</sup>, The Robbers Cave Experiment<sup>4</sup>

Cade Metz at *The Register* recently<sup>5</sup> alleged<sup>6</sup> 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<sup>7</sup>.)

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<sup>8</sup> 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.

The ingroup-outgroup dichotomy<sup>9</sup> is part of ordinary human nature. So are happy death spirals<sup>10</sup> and spirals of hate<sup>11</sup>. 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. Ev-

---

1. [http://lesswrong.com/lw/lv/every\\_cause\\_wants\\_to\\_be\\_a\\_cult/](http://lesswrong.com/lw/lv/every_cause_wants_to_be_a_cult/)

2. Page 326, ‘Correspondence Bias’.

3. Page 384, ‘Affective Death Spirals’.

4. Page 332, ‘The Robbers Cave Experiment’.

5. [http://www.theregister.co.uk/2007/12/04/wikipedia\\_secret\\_mailing/](http://www.theregister.co.uk/2007/12/04/wikipedia_secret_mailing/)

6. [http://www.theregister.co.uk/2007/12/06/wikipedia\\_and\\_overstock/](http://www.theregister.co.uk/2007/12/06/wikipedia_and_overstock/)

7. <http://postbiota.org/pipermail/tt/2007-December/001947.html>

8. Page 326, ‘Correspondence Bias’.

9. Page 332, ‘The Robbers Cave Experiment’.

10. Page 384, ‘Affective Death Spirals’.

11. Page 400, ‘When None Dare Urge Restraint’.

erything 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<sup>12</sup>. 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 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

---

<sup>12</sup>. Page 374, ‘The Halo Effect’.

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<sup>13</sup> 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<sup>14</sup> such<sup>15</sup> 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 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. Page 465, ‘How to Seem (and Be) Deep’.

14. Page 386, ‘Resist the Happy Death Spiral’.

15. Page 640, ‘The Meditation on Curiosity’.

## 14. Guardians of the Truth<sup>1</sup>

**Followup to:** Tsuyoku Naritai<sup>2</sup>, Reversed Stupidity is not Intelligence<sup>3</sup>

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<sup>4</sup>?”

Reversed stupidity is not intelligence<sup>5</sup>: “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 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

1. [http://lesswrong.com/lw/lz/guardians\\_of\\_the\\_truth/](http://lesswrong.com/lw/lz/guardians_of_the_truth/)

2. [http://lesswrong.com/lw/b8/tsuyoku\\_naritai\\_i\\_want\\_to\\_become\\_stronger/](http://lesswrong.com/lw/b8/tsuyoku_naritai_i_want_to_become_stronger/)

3. Page 335, ‘Reversed Stupidity Is Not Intelligence’.

4. <http://yudkowsky.net/bayes/truth.html>

5. Page 335, ‘Reversed Stupidity Is Not Intelligence’.

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<sup>6</sup>

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 “ $2 + 2 = 4$ ”.

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.

---

6. <http://books.google.com/books?id=LDOCZPyg2M&q=PA76&vq=xii&dq=%22chuang+tzu%22+deeds+history&sig=govZRcDAVbAQdvi2c>

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<sup>7</sup>. 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 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.

---

7. [http://lesswrong.com/lw/gv/outside\\_the\\_laboratory/](http://lesswrong.com/lw/gv/outside_the_laboratory/)

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<sup>8</sup> 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 406, ‘Every Cause Wants To Be A Cult’.



## 15. Guardians of the Gene Pool<sup>1</sup>

### Followup to: Guardians of the Truth<sup>2</sup>

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/](http://lesswrong.com/lw/mo/guardians_of_the_gene_pool/)

2. Page 409, ‘Guardians of the Truth’.

## 16. Guardians of Ayn Rand<sup>1</sup>

**Followup to:** Every Cause Wants To Be A Cult<sup>2</sup>, Guardians of the Truth<sup>3</sup>

“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<sup>4</sup>”

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<sup>5</sup>. Her ultimate fictional hero, John Galt, was <SPOILER>  
տօլես ու երօն մի շուրջ թօւառաւուց:  
բուժ բույս ու ճակա ի ու բուժ առաջ գնուց երաց երաց այլ ք  
և ա ՏԿԱՆԻՑ արօ լաւուց և սամ լուս օվ ըստ եւսարից ըստ ման:  
</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 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

1. [http://lesswrong.com/lw/m1/guardians\\_of\\_ayn\\_rand/](http://lesswrong.com/lw/m1/guardians_of_ayn_rand/)

2. Page 406, ‘Every Cause Wants To Be A Cult’.

3. Page 409, ‘Guardians of the Truth’.

4. [http://www.2think.org/o2\\_2\\_sbe.shtml](http://www.2think.org/o2_2_sbe.shtml)

5. [http://lesswrong.com/lw/ky/fake\\_morality/](http://lesswrong.com/lw/ky/fake_morality/)

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<sup>6</sup>: "[Barbara] Branden recalled an evening when a friend of Rand's remarked that he enjoyed the music of Richard

---

6. [http://www.2think.org/o2\\_2\\_sbe.shtml](http://www.2think.org/o2_2_sbe.shtml)

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<sup>7</sup>.”

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<sup>8</sup> somehow leap to mind.

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<sup>9</sup> 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<sup>10</sup>.

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<sup>11</sup>), and she enjoyed it too

7. Page 625, ‘Feeling Rational’.

8. Page 538, ‘Fake Optimization Criteria’.

9. Page 326, ‘Correspondence Bias’.

10. Page 406, ‘Every Cause Wants To Be A Cult’.

11. Page 384, ‘Affective Death Spirals’.

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.<sup>12</sup>) Ever after, Rand's grip over 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<sup>13</sup>.

Let that be a lesson to all of us: Praising<sup>14</sup> "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.

---

12. Page 396, 'Evaporative Cooling of Group Beliefs'.

13. [http://lesswrong.com/lw/hs/think\\_like\\_reality/](http://lesswrong.com/lw/hs/think_like_reality/)

14. Page 144, 'Science as Attire'.

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

### **工程师**

; 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<sup>15</sup>.

Now me, y’know, I admire Francis Bacon’s audacity<sup>16</sup>, 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<sup>17</sup>, 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.

---

15. [http://lesswrong.com/lw/b8/tsuyoku\\_naritai\\_i\\_want\\_to\\_become\\_stronger/](http://lesswrong.com/lw/b8/tsuyoku_naritai_i_want_to_become_stronger/)

16. Page 386, ‘Resist the Happy Death Spiral’.

17. <http://singinst.org/AIRisk.pdf>

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.

## 17. The Litany Against Gurus<sup>1</sup>

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.

---

<sup>1</sup>. [http://lesswrong.com/lw/m2/the\\_litany\\_against\\_gurus/](http://lesswrong.com/lw/m2/the_litany_against_gurus/)

## 18. Two Cult Koans<sup>1</sup>

**Followup to:** Every Cause Wants To Be A Cult<sup>2</sup>

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 reassurance. If you had been truly curious<sup>3</sup>, not knowing one way or the other, you would have thought of ways to resolve your doubts<sup>4</sup>. 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

---

1. [http://lesswrong.com/lw/m4/two\\_cult\\_koans/](http://lesswrong.com/lw/m4/two_cult_koans/)

2. Page 406, ‘Every Cause Wants To Be A Cult’.

3. Page 640, ‘The Meditation on Curiosity’.

4. Page 636, ‘The Proper Use of Doubt’.

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 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.

## 19. Asch's Conformity Experiment<sup>1</sup>

<sup>2</sup> 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 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

1. [http://lesswrong.com/lw/m9/aschs\\_conformity\\_experiment/](http://lesswrong.com/lw/m9/aschs_conformity_experiment/)

2. <http://scienceaid.co.uk/psychology/social/images/asch.png>

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<sup>3</sup> 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 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

---

3. Page 126, ‘Your Strength as a Rationalist’.

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 Task4. *Psychological Bulletin*, **119**, 111-137.

---

4. [http://www.radford.edu/~jaspelme/\\_private/gradsoc\\_articles/individualism\\_collectivism/conformity%20and%20culture.pdf](http://www.radford.edu/~jaspelme/_private/gradsoc_articles/individualism_collectivism/conformity%20and%20culture.pdf)

## 20. Lonely Dissent<sup>1</sup>

**Followup to:** The Modesty Argument<sup>2</sup>, The “Outside the Box” Box<sup>3</sup>, Asch’s Conformity Experiment<sup>4</sup>

Asch’s conformity experiment<sup>5</sup> 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<sup>6</sup>. 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 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 re-

---

1. [http://lesswrong.com/lw\(mb/lonely\\_dissent/](http://lesswrong.com/lw(mb/lonely_dissent/)

2. [http://lesswrong.com/lw/gr/the\\_modesty\\_argument/](http://lesswrong.com/lw/gr/the_modesty_argument/)

3. Page 454, ‘The “Outside the Box” Box’.

4. Page 424, ‘Asch’s Conformity Experiment’.

5. Page 424, ‘Asch’s Conformity Experiment’.

6. Page 454, ‘The “Outside the Box” Box’.

spect 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<sup>7</sup>. There are other cryonicists 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<sup>8</sup>: 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

7. <http://www.alcor.org/>

8. [http://lesswrong.com/lw/lx/evolutionary\\_psychology/](http://lesswrong.com/lw/lx/evolutionary_psychology/)

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<sup>9</sup> 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 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<sup>10</sup>. 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 over-

---

9. [http://lesswrong.com/lw/lx/evolutionary\\_psychology/](http://lesswrong.com/lw/lx/evolutionary_psychology/)

10. [http://www.overcomingbias.com/2007/06/against\\_free\\_tb.html](http://www.overcomingbias.com/2007/06/against_free_tb.html)

whelmingly 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.

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<sup>11</sup>. It's not a *virtue*, you see—just a bias either way.

---

11. [http://www.overcomingbias.com/2007/06/against\\_free\\_tb.html](http://www.overcomingbias.com/2007/06/against_free_tb.html)

## 21. Cultish Countercultishness<sup>1</sup>

**Followup to:** Every Cause Wants To Be A Cult<sup>2</sup>, Lonely Dissent<sup>3</sup>

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<sup>4</sup>, 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 like group polarization<sup>5</sup>, happy death spirals<sup>6</sup>, uncriticality<sup>7</sup>, and evaporative cooling<sup>8</sup>, all of which seem to feed on each other. When these failures swirl together and meet, they combine to form a Super-Failure stupider

---

1. [http://lesswrong.com/lw/md/cultish\\_countercultishness/](http://lesswrong.com/lw/md/cultish_countercultishness/)

2. Page 406, 'Every Cause Wants To Be A Cult'.

3. Page 428, 'Lonely Dissent'.

4. <http://www.prem-ravat-talk.org/forum/uploads/CultCharacteristics.htm>

5. Page 332, 'The Robbers Cave Experiment'.

6. Page 384, 'Affective Death Spirals'.

7. Page 392, 'Uncritical Supercriticality'.

8. Page 396, 'Evaporative Cooling of Group Beliefs'.

than any of the parts, like Voltron<sup>9</sup>. 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 semidog.

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<sup>10</sup> and see a positive halo effect<sup>11</sup> around their Favorite Thing Ever—which could be Objectivism<sup>12</sup>, or vegetarianism, or neural networks<sup>13</sup>—you cannot, *from the evidence gathered so far*, deduce whether they have achieved uncriticality<sup>14</sup>. You cannot deduce whether their 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!"

---

9. [http://www.youtube.com/watch?v=tZZv5Z2Iz\\_s](http://www.youtube.com/watch?v=tZZv5Z2Iz_s)

10. Page 332, 'The Robbers Cave Experiment'.

11. Page 374, 'The Halo Effect'.

12. Page 415, 'Guardians of Ayn Rand'.

13. [http://thedailywtf.com/Articles/No,\\_We\\_Need\\_a\\_Neural\\_Network.aspx](http://thedailywtf.com/Articles/No,_We_Need_a_Neural_Network.aspx)

14. Page 392, 'Uncritical Supercriticality'.

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<sup>15</sup>. 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<sup>16</sup>. 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<sup>17</sup>, this<sup>18</sup> and this<sup>19</sup>). 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 questions of rationality. You will not be genuinely curious<sup>20</sup> or think of ways to fulfill your doubts<sup>21</sup>. 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<sup>22</sup>, whether the cause itself is wise or foolish. The ingroup-outgroup dichotomy<sup>23</sup> etc. are part of human nature, not a special curse<sup>24</sup> of mutants<sup>25</sup>. Rationality is the exception, not the rule. You have to put forth a constant effort to maintain rationality against the natural slide into entropy.

---

15. Page 335, 'Reversed Stupidity Is Not Intelligence'.

16. Page 343, 'Hug the Query'.

17. <http://yudkowsky.net/virtues/>

18. Page 83, 'Why truth? And...?'.

19. Page 422, 'Two Cult Koans'.

20. Page 640, 'The Meditation on Curiosity'.

21. Page 636, 'The Proper Use of Doubt'.

22. Page 406, 'Every Cause Wants To Be A Cult'.

23. Page 332, 'The Robbers Cave Experiment'.

24. Page 326, 'Correspondence Bias'.

25. Page 329, 'Are Your Enemies Innately Evil?'.

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<sup>26</sup>, and will have to pump against entropy to avoid an affective death spiral<sup>27</sup>. 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<sup>28</sup> 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, 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<sup>29</sup> and happy death spirals<sup>30</sup> can happen in political discussion, in mainstream religions, in sports fandom. If the *nervousness* came from fear of *rationality errors*, people would ask “This

26. Page 374, ‘The Halo Effect’.

27. Page 384, ‘Affective Death Spirals’.

28. Page 374, ‘The Halo Effect’.

29. Page 332, ‘The Robbers Cave Experiment’.

30. Page 384, ‘Affective Death Spirals’.

isn't an ingroup-outgroup dichotomy<sup>31</sup>, 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 cryonics 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<sup>32</sup>?

I suspect that the *nervousness* is not the fear of believing falsely, or the fear of physical harm. It is the fear of lonely dissent<sup>33</sup>. The nervous feeling that subjects get in Asch's conformity experiment<sup>34</sup>, 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.

---

31. Page 332, 'The Robbers Cave Experiment'.

32. <http://www.alcor.org/>

33. Page 428, 'Lonely Dissent'.

34. Page 424, 'Asch's Conformity Experiment'.

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<sup>35</sup> and chanting<sup>36</sup>. 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 everyone 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<sup>37</sup> and evaporative cooling<sup>38</sup> 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<sup>39</sup>. But that makes lofty goals a risk factor, not a disease. Some goals are genuinely worth pursuing<sup>40</sup>.

On the other hand, I see no legitimate reason for sleep deprivation or threatening dissenters with beating, full stop<sup>41</sup>. When a group does this, then whether you call it "cult" or "not-cult", you have directly answered<sup>42</sup> 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<sup>43</sup>, 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.

35. Page 422, 'Two Cult Koans'.

36. Page 349, 'The Litany Against Gurus'.

37. Page 332, 'The Robbers Cave Experiment'.

38. Page 396, 'Evaporative Cooling of Group Beliefs'.

39. Page 384, 'Affective Death Spirals'.

40. Page 386, 'Resist the Happy Death Spiral'.

41. Page 392, 'Uncritical Supercriticality'.

42. Page 343, 'Hug the Query'.

43. Page 424, 'Asch's Conformity Experiment'.

The one wishes there was some way they could be *sure* that they weren't in a "cult". Some definite, crushing rejoinder to 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<sup>44</sup> 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<sup>45</sup>, 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<sup>46</sup> (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. 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*.

---

44. Page 636, 'The Proper Use of Doubt'.

45. <http://singinst.org/AIRisk.pdf>

46. Page 392, 'Uncritical Supercriticality'.

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<sup>47</sup> 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<sup>48</sup>, 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

---

47. Page 422, ‘Two Cult Koans’.

48. Page 343, ‘Hug the Query’.

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<sup>49</sup>. 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.)

---

49. [http://lesswrong.com/lw/kg/expecting\\_short\\_inferential\\_distances/](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...*



## I. Anchoring and Adjustment<sup>1</sup>

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!*

$$1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8$$

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 “ $1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8$ ” made a median estimate of 512, while students shown “ $8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$ ” 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.

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.

---

<sup>1</sup>. [http://lesswrong.com/lw/j7/anchoring\\_and\\_adjustment/](http://lesswrong.com/lw/j7/anchoring_and_adjustment/)

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<sup>1</sup>

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—or of the following letter-string: “**GREEN**”

In Mussweiler and Strack (2000), subjects were asked the anchoring question<sup>2</sup>: “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 estimates and decisions. In the field of heuristics and biases, this more general phenomenon is known as *contamination*. (Chapman and Johnson 2002.)

---

1. [http://lesswrong.com/lw/k3/priming\\_and\\_contamination/](http://lesswrong.com/lw/k3/priming_and_contamination/)

2. Page 443, ‘Anchoring and Adjustment’.

Early research in heuristics and biases discovered anchoring effects<sup>3</sup>, 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<sup>4</sup> 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<sup>5</sup>. 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<sup>6</sup> another<sup>7</sup> of<sup>8</sup> the<sup>9</sup> thousand<sup>10</sup> faces<sup>11</sup> of<sup>12</sup> confirmation<sup>13</sup> bias<sup>14</sup>. 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

3. Page 443, ‘Anchoring and Adjustment’.

4. [http://lesswrong.com/lw/j7/anchoring\\_and\\_adjustment/eza](http://lesswrong.com/lw/j7/anchoring_and_adjustment/eza)

5. [http://en.wikipedia.org/wiki/Bias\\_blind\\_spot](http://en.wikipedia.org/wiki/Bias_blind_spot)

6. Page 164, ‘Positive Bias: Look Into the Dark’.

7. Page 505, ‘Knowing About Biases Can Hurt People’.

8. Page 589, ‘The Third Alternative’.

9. Page 133, ‘Hindsight bias’.

10. Page 469, ‘We Change Our Minds Less Often Than We Think’.

11. Page 147, ‘Fake Causality’.

12. Page 511, ‘One Argument Against An Army’.

13. Page 519, ‘Rationalization’.

14. Page 516, ‘What Evidence Filtered Evidence?’.

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<sup>15</sup>, for we change our minds less often than we think<sup>16</sup>...

---

Chapman, G.B. and Johnson, E.J. 2002. Incorporating the irrelevant: Anchors in judgments of belief and value<sup>17</sup>. 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.

---

15. Page 513, 'The Bottom Line'.

16. Page 469, '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?<sup>1</sup>

Some early experiments on anchoring<sup>2</sup> and adjustment<sup>3</sup> 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<sup>4</sup>.

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<sup>5</sup>. 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 should not cause them to remember true statements as being false.

Gilbert, Krull, and Malone<sup>6</sup> (1990) bears out this result, showing that, among subjects presented with novel statements labeled TRUE or FALSE, distraction had no effect on identifying true

---

1. [http://lesswrong.com/lw/k4/do\\_we\\_believe\\_everything\\_were\\_told/](http://lesswrong.com/lw/k4/do_we_believe_everything_were_told/)

2. Page 445, ‘Priming and Contamination’.

3. Page 443, ‘Anchoring and Adjustment’.

4. Page 445, ‘Priming and Contamination’.

5. <http://www.overcomingbias.com/2007/10/what-evidence-i.html>

6. <http://www.wjh.harvard.edu/%07Edtg/Gilbert%20et%20al%20%28UNBELIEVING%29.pdf>

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<sup>7</sup> (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 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.

---

<sup>7</sup>. [http://www.wjh.harvard.edu/~dtg/Gilbert%20et%20al%20\(EVERYTHING%20YOU%20READ\).pdf](http://www.wjh.harvard.edu/~dtg/Gilbert%20et%20al%20(EVERYTHING%20YOU%20READ).pdf)

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.<sup>8</sup> *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.<sup>9</sup> *Journal of Personality and Social Psychology*, **65**(2), 221-233.

---

8. <http://www.wjh.harvard.edu/%07Edtg/Gilbert%20et%20al%20%28UNBELIEVING%29.pdf>

9. [http://www.wjh.harvard.edu/~dtg/Gilbert%20et%20al%20\(EVERYTHING%20YOU%20READ\).pdf](http://www.wjh.harvard.edu/~dtg/Gilbert%20et%20al%20(EVERYTHING%20YOU%20READ).pdf)

## 4. Cached Thoughts<sup>1</sup>

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 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

---

<sup>1</sup>. [http://lesswrong.com/lw/k5/cached\\_thoughts/](http://lesswrong.com/lw/k5/cached_thoughts/)

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<sup>2</sup>, 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 evidence 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,

---

2. [http://lesswrong.com/lw/i8/religions\\_claim\\_to\\_be\\_nondisprovable/](http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/)

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<sup>3</sup>, 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 625, ‘Feeling Rational’.

## 5. The “Outside the Box” Box<sup>1</sup>

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<sup>2</sup> 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<sup>3</sup> 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<sup>4</sup>:

“Revolution has already been televised. Revolution has been \*merchandised\*. Revolution is a commodity, a

---

1. [http://lesswrong.com/lw/k6/the\\_outside\\_the\\_box\\_box/](http://lesswrong.com/lw/k6/the_outside_the_box_box/)

2. Page 451, ‘Cached Thoughts’.

3. <http://scottaaronson.com/blog/?p=87#comment-2150>

4. <http://journalism.berkeley.edu/projects/biplog/archive/000748.html>

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<sup>5</sup>. 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 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 mouse-trap!”, 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

---

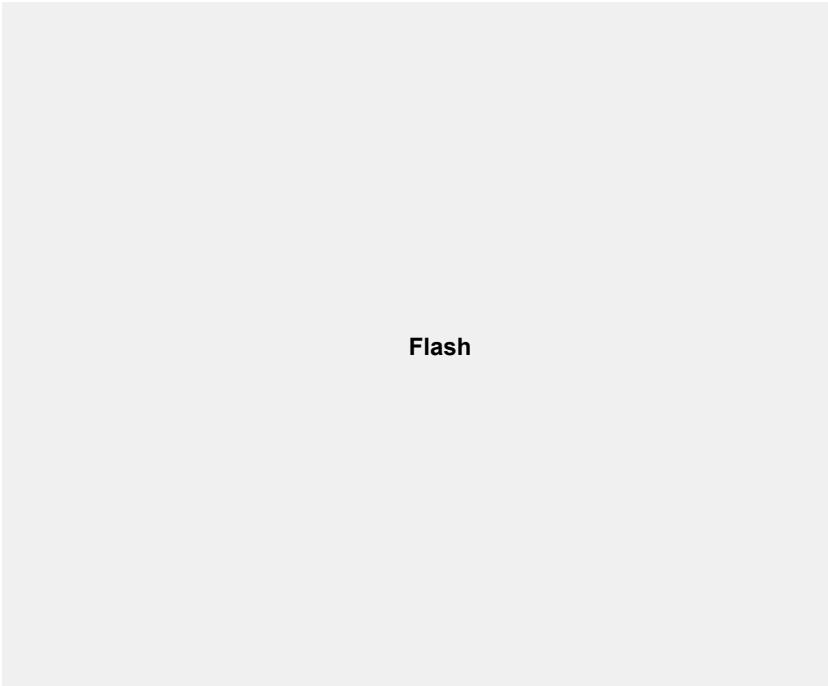
<sup>5</sup>. Page 451, ‘Cached Thoughts’.

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<sup>6</sup> said it best of all:



**Flash**

---

6. <http://www.youtube.com/watch?v=LQgq3eo3EBQ>

## 6. Original Seeing<sup>1</sup>

**Followup to:** Cached Thoughts<sup>2</sup>, The Virtue of Narrowness<sup>3</sup>

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.

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

---

1. [http://lesswrong.com/lw/k7/original\\_seeing/](http://lesswrong.com/lw/k7/original_seeing/)

2. Page 451, 'Cached Thoughts'.

3. Page 122, 'The Virtue of Narrowness'.

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<sup>1</sup>

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<sup>2</sup>:

"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 successfully repel an invasion of monsters or robot warriors, it wouldn't be much fun to watch."

So there are specific distortions<sup>3</sup> in fiction. But trying to correct for these specific distortions is not enough. A story is *never* a

---

1. [http://lesswrong.com/lw/k9/the\\_logical\\_fallacy\\_of\\_generalization\\_from/](http://lesswrong.com/lw/k9/the_logical_fallacy_of_generalization_from/)

2. <http://www.nickbostrom.com/existential/risks.html>

3. [http://www.overcomingbias.com/2006/12/biases\\_of\\_scien.html](http://www.overcomingbias.com/2006/12/biases_of_scien.html)

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 Merklethud 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<sup>4</sup> that lend flesh to the story, from Wifflefoofer's appropriately futuristic attitudes toward feminism, down to the color of her earrings.

Then all these burdensome details<sup>5</sup> and questionable assumptions are wrapped up and given a short label<sup>6</sup>, creating the illusion that they are a single package<sup>7</sup>.

On problems with large answer spaces, the greatest difficulty is not *verifying* the correct answer but simply locating it in answer space<sup>8</sup> to begin with. If someone starts out by asking whether or

4. [http://lesswrong.com/lw/jk/burdensome\\_details/](http://lesswrong.com/lw/jk/burdensome_details/)

5. [http://lesswrong.com/lw/jk/burdensome\\_details/](http://lesswrong.com/lw/jk/burdensome_details/)

6. Page 95, 'Occam's Razor'.

7. [http://en.wikipedia.org/wiki/Package-deal\\_fallacy](http://en.wikipedia.org/wiki/Package-deal_fallacy)

8. [http://wiki.lesswrong.com/wiki/Locate\\_the\\_hypothesis](http://wiki.lesswrong.com/wiki/Locate_the_hypothesis)

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<sup>9</sup> and widen confidence intervals<sup>10</sup>, 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<sup>11</sup>.

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 fingers. 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<sup>12</sup> of smarter-than-human intelligence and the argument for its unpredictability<sup>13</sup>; 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

9. [http://lesswrong.com/lw/j1/stranger\\_than\\_history/](http://lesswrong.com/lw/j1/stranger_than_history/)

10. [http://lesswrong.com/lw/j6/why\\_is\\_the\\_future\\_so\\_absurd/](http://lesswrong.com/lw/j6/why_is_the_future_so_absurd/)

11. [http://wiki.lesswrong.com/wiki/Privileging\\_the\\_hypothesis](http://wiki.lesswrong.com/wiki/Privileging_the_hypothesis)

12. <http://www.singinst.org/blog/2007/07/10/the-power-of-intelligence/>

13. <http://www.singinst.org/blog/2007/09/30/three-major-singularity-schools/>

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<sup>14</sup> us and make compatible thoughts more available<sup>15</sup>, 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 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<sup>16</sup> 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<sup>17</sup>. 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

<sup>14</sup>. Page 445, ‘Priming and Contamination’.

<sup>15</sup>. <http://lesswrong.com/lw/j5/availability/>

<sup>16</sup>. Page 448, ‘Do We Believe Everything We’re Told?’.

<sup>17</sup>. <http://lesswrong.com/lw/j5/availability/>

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 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<sup>18</sup>.)

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."

---

18. Page 147, 'Fake Causality'.

As George Orwell said<sup>19</sup> of cliches:

“What is above all needed is to let the meaning choose the word, and not the other way around... 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<sup>20</sup>:

“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<sup>21</sup>, Contamination<sup>22</sup>, Availability<sup>23</sup>, Cached Thoughts<sup>24</sup>, Do We Believe Everything We're Told?<sup>25</sup>, Einstein's Arrogance<sup>26</sup>, Burdensome details<sup>27</sup>

19. Page 345, ‘Rationality and the English Language’.

20. Page 457, ‘Original Seeing’.

21. Page 443, ‘Anchoring and Adjustment’.

22. Page 445, ‘Priming and Contamination’.

23. <http://lesswrong.com/lw/j5/availability/>

24. Page 451, ‘Cached Thoughts’.

25. Page 448, ‘Do We Believe Everything We're Told?’.

26. [http://lesswrong.com/lw/jo/einsteins\\_arrogance/](http://lesswrong.com/lw/jo/einsteins_arrogance/)

27. [http://lesswrong.com/lw/jk/burdensome\\_details/](http://lesswrong.com/lw/jk/burdensome_details/)

## 8. How to Seem (and Be) Deep<sup>1</sup>

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<sup>2</sup>, 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<sup>3</sup>.

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 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

---

1. [http://lesswrong.com/lw/k8/how\\_to\\_seem\\_and\\_be\\_deep/](http://lesswrong.com/lw/k8/how_to_seem_and_be_deep/)

2. Page 457, 'Original Seeing'.

3. <http://www.singinst.org/blog/2007/06/16/transhumanism-as-simplified-humanism/>

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<sup>4</sup> 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<sup>5</sup>; they will not have heard anything they could not have thought of for themselves. One might call it belief in wisdom<sup>6</sup>—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<sup>7</sup> as it were, because they're trying to seem Deeply Wise instead of optimizing<sup>8</sup>.

How much thinking did I need to do, in the course of seeming deep? Human brains only run at 100Hz and I responded in real-time, 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,

---

4. Page 451, 'Cached Thoughts'.

5. Page 454, 'The "Outside the Box" Box'.

6. Page 108, 'Belief in Belief'.

7. Page 457, 'Original Seeing'.

8. Page 454, 'The "Outside the Box" Box'.

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<sup>9</sup>. 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<sup>10</sup> depicted as repositories of deep wisdom and/or mystical secrets. (And sometimes not<sup>11</sup>.)

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 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.

---

9. <http://www.singinst.org/blog/2007/06/16/transhumanism-as-simplified-humanism/>

10. <http://tv tropes.org/pmwiki/pmwiki.php>Main/JesusTaboo>

11. <http://tv tropes.org/pmwiki/pmwiki.php>Main/CreepyCoolCrosses>

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<sup>12</sup>. It does mean seeing for yourself<sup>13</sup>, rather than letting your brain complete the pattern<sup>14</sup>. If you don’t stop at the first answer<sup>15</sup>, and cast out replies that seem vaguely unsatisfactory<sup>16</sup>, in time your thoughts will form a coherent whole, flowing from the single source of yourself, rather than being fragmentary repetitions<sup>17</sup> of other people’s conclusions<sup>18</sup>.

---

12. Page 454, ‘The “Outside the Box” Box’.

13. Page 457, ‘Original Seeing’.

14. Page 451, ‘Cached Thoughts’.

15. Page 589, ‘The Third Alternative’.

16. Page 126, ‘Your Strength as a Rationalist’.

17. Page 457, ‘Original Seeing’.

18. Page 451, ‘Cached Thoughts’.

## 9. We Change Our Minds Less Often Than We Think<sup>1</sup>

“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<sup>2</sup> 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.

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.

---

1. [http://lesswrong.com/lw/jx/we\\_change\\_our\\_minds\\_less\\_often\\_than\\_we\\_think/](http://lesswrong.com/lw/jx/we_change_our_minds_less_often_than_we_think/)

2. Page 513, ‘The Bottom Line’.

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<sup>3</sup>, in your heuristic estimate of your competence<sup>4</sup>?

Between hindsight bias<sup>5</sup>, fake causality<sup>6</sup>, positive bias<sup>7</sup>, anchoring<sup>8</sup>/priming, et cetera et cetera, and above all the dreaded confirmation bias<sup>9</sup>, 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](http://en.wikipedia.org/wiki/Lake_Wobegon_effect)

5. Page 133, 'Hindsight bias'.

6. Page 147, 'Fake Causality'.

7. Page 164, 'Positive Bias: Look Into the Dark'.

8. Page 443, 'Anchoring and Adjustment'.

9. Page 505, 'Knowing About Biases Can Hurt People'.

## 10. Hold Off On Proposing Solutions<sup>1</sup>

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 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

---

<sup>1</sup>. [http://lesswrong.com/lw/ka/hold\\_off\\_on\\_proposing\\_solutions/](http://lesswrong.com/lw/ka/hold_off_on_proposing_solutions/)

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<sup>2</sup>, loosely speaking—why, *that* problem is so incredibly difficult that an actual majority resolve the whole issue within 15 seconds. *Give me a break.*

**(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<sup>3</sup>, 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<sup>4</sup> above. If you make your decision very early on, it will, in fact, be

---

2. <http://singinst.org/AIRisk.pdf>

3. Page 513, ‘The Bottom Line’.

4. Page 519, ‘Rationalization’.

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<sup>5</sup>: 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<sup>6</sup> 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 *bold 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.

Even half a minute would be an improvement over half a second.

---

5. Page 469, 'We Change Our Minds Less Often Than We Think'.

6. Page 650, 'No One Can Exempt You From Rationality's Laws'.

## 11. Asch's Conformity Experiment<sup>1</sup>

<sup>2</sup> 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 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

1. [http://lesswrong.com/lw/m9/aschs\\_conformity\\_experiment/](http://lesswrong.com/lw/m9/aschs_conformity_experiment/)

2. <http://scienceaid.co.uk/psychology/social/images/asch.png>

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<sup>3</sup> 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 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

---

3. Page 126, ‘Your Strength as a Rationalist’.

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 Task4. *Psychological Bulletin*, **119**, 111-137.

---

4. [http://www.radford.edu/~jaspelme/\\_private/gradsoc\\_articles/individualism\\_collectivism/conformity%20and%20culture.pdf](http://www.radford.edu/~jaspelme/_private/gradsoc_articles/individualism_collectivism/conformity%20and%20culture.pdf)

## 12. On Expressing Your Concerns<sup>1</sup>

### Followup to: Asch's Conformity Experiment<sup>2</sup>

The scary thing about Asch's conformity experiments<sup>3</sup> 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<sup>4</sup> 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 optimistic 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.

---

1. [http://lesswrong.com/lw/ma/on\\_expressing\\_your\\_concerns/](http://lesswrong.com/lw/ma/on_expressing_your_concerns/)

2. Page 424, ‘Asch’s Conformity Experiment’.

3. Page 424, ‘Asch’s Conformity Experiment’.

4. [http://lesswrong.com/lw/jg/planning\\_fallacy/](http://lesswrong.com/lw/jg/planning_fallacy/)

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<sup>5</sup> disagreement shows *someone* must be irrational<sup>6</sup>. 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.”<sup>7</sup>

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 know, I know, you know I know, etc. Hanson's post or paper on “We Can't Foresee to Disagree”<sup>8</sup> 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

<sup>5</sup>. Page 625, ‘Feeling Rational’.

<sup>6</sup>. [http://lesswrong.com/lw/gr/the\\_modesty\\_argument/](http://lesswrong.com/lw/gr/the_modesty_argument/)

<sup>7</sup>. [http://www.overcomingbias.com/2006/12/you\\_are\\_never\\_e.html](http://www.overcomingbias.com/2006/12/you_are_never_e.html)

<sup>8</sup>. [http://www.overcomingbias.com/2007/01/we\\_cant\\_foresee.html](http://www.overcomingbias.com/2007/01/we_cant_foresee.html)

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.

## 13. Lonely Dissent<sup>1</sup>

**Followup to:** The Modesty Argument<sup>2</sup>, The “Outside the Box” Box<sup>3</sup>, Asch’s Conformity Experiment<sup>4</sup>

Asch’s conformity experiment<sup>5</sup> 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<sup>6</sup>. 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 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 re-

---

1. [http://lesswrong.com/lw\(mb/lonely\\_dissent/](http://lesswrong.com/lw(mb/lonely_dissent/)

2. [http://lesswrong.com/lw/gr/the\\_modesty\\_argument/](http://lesswrong.com/lw/gr/the_modesty_argument/)

3. Page 454, ‘The “Outside the Box” Box’.

4. Page 424, ‘Asch’s Conformity Experiment’.

5. Page 424, ‘Asch’s Conformity Experiment’.

6. Page 454, ‘The “Outside the Box” Box’.

spect 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<sup>7</sup>. There are other cryonicists 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<sup>8</sup>: 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

7. <http://www.alcor.org/>

8. [http://lesswrong.com/lw/lx/evolutionary\\_psychology/](http://lesswrong.com/lw/lx/evolutionary_psychology/)

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<sup>9</sup> 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 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<sup>10</sup>. 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 over-

---

9. [http://lesswrong.com/lw/lx/evolutionary\\_psychology/](http://lesswrong.com/lw/lx/evolutionary_psychology/)

10. [http://www.overcomingbias.com/2007/06/against\\_free\\_tb.html](http://www.overcomingbias.com/2007/06/against_free_tb.html)

whelmingly 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.

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<sup>ii</sup>. It's not a *virtue*, you see—just a bias either way.

---

ii. [http://www.overcomingbias.com/2007/06/against\\_free\\_tb.html](http://www.overcomingbias.com/2007/06/against_free_tb.html)

## 14. The Genetic Fallacy<sup>1</sup>

In lists<sup>2</sup> of<sup>3</sup> logical<sup>4</sup> fallacies<sup>5</sup>, 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<sup>6</sup> 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<sup>7</sup>. Genetic accusations have a force among humans that they would not have among ideal Bayesians.

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

1. [http://lesswrong.com/lw/s3/the\\_genetic\\_fallacy/](http://lesswrong.com/lw/s3/the_genetic_fallacy/)

2. [http://www.nizkor.org/features/fallacies/genetic\\_fallacy.html](http://www.nizkor.org/features/fallacies/genetic_fallacy.html)

3. [http://en.wikipedia.org/wiki/List\\_of\\_fallacies](http://en.wikipedia.org/wiki/List_of_fallacies)

4. [http://atheism.about.com/library/FAQs/skepticism/blfaq\\_fall\\_genetic.htm](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 469, ‘We Change Our Minds Less Often Than We Think’.

level) the idea that the Bible was written by the hand of God; and who nonetheless think that the Bible contains indispensable ethical wisdom<sup>8</sup>. 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<sup>9</sup>; 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<sup>10</sup> cached<sup>11</sup> 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-cut evidence is what Science<sup>12</sup> 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<sup>13</sup> 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.

---

8. [http://lesswrong.com/lw/i8/religions\\_claim\\_to\\_be\\_nondisprovable/](http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/)

9. Page 335, ‘Reversed Stupidity Is Not Intelligence’.

10. Page 511, ‘One Argument Against An Army’.

11. Page 451, ‘Cached Thoughts’.

12. [http://lesswrong.com/lw/in/scientific\\_evidence\\_legal\\_evidence\\_rational/](http://lesswrong.com/lw/in/scientific_evidence_legal_evidence_rational/)

13. [http://lesswrong.com/lw/jo/einsteins\\_arrogance/](http://lesswrong.com/lw/jo/einsteins_arrogance/)

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...
- Be extremely suspicious if you find that you still believe the early suggestions of a source you later rejected.

**Added:** Hal Finney suggests<sup>14</sup> that we should call it “the genetic heuristic”.

---

<sup>14</sup>. <http://www.overcomingbias.com/2008/07/genetic-fallacy.html#comment-121978890>



## **Noticing Confusion**

*(Heavy overlap with Mysterious Answers to  
Mysterious Questions.)*



## I. Your Strength as a Rationalist<sup>1</sup>

(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 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...

---

<sup>1</sup>. [http://lesswrong.com/lw/if/your\\_strength\\_as\\_a\\_rationalist/](http://lesswrong.com/lw/if/your_strength_as_a_rationalist/)

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<sup>2</sup> than a published journal article. Alas, belief is easier than disbelief; we believe instinctively, but disbelief requires a conscious effort<sup>3</sup>.

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<sup>4</sup>.

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

---

2. <http://www.overcomingbias.com/2007/08/truth-bias.html>

3. <http://www.wjh.harvard.edu/~o7Edtg/>

*Gilbert%20et%20al%20%28EVERYTHING%20YOU%20READ%29.pdf*

4. Page 105, 'Making Beliefs Pay Rent (in Anticipated Experiences)'.

## 2. Absence of Evidence Is Evidence of Absence<sup>1</sup>

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<sup>2</sup>. 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 is still higher that the *absence* of a Fifth Column would perform an absence of sabotage.

Let E stand for the observation of sabotage, H<sub>1</sub> for the hypothesis of a Japanese-American Fifth Column, and H<sub>2</sub> for the hypothesis that no Fifth Column exists. Whatever the likelihood

1. [http://lesswrong.com/lw/ib/absence\\_of\\_evidence\\_is\\_evidence\\_of\\_absence/](http://lesswrong.com/lw/ib/absence_of_evidence_is_evidence_of_absence/)

2. Page 28, 'An Intuitive Explanation of Bayes' Theorem'.

that a Fifth Column would do no sabotage, the probability  $P(E|H_1)$ , it cannot be as large as the likelihood that no Fifth Column does no sabotage, the probability  $P(E|H_2)$ . 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,  $A \rightarrow B$ , "A implies B", is not equivalent to  $\neg A \rightarrow \neg B$ , "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  $P(H| \neg E) < P(H)$ , "failure to observe E decreases the probability of H".  $P(H)$  is a weighted mix of  $P(H|E)$  and  $P(H| \neg E)$ , and necessarily lies between the two. If any of this sounds at all confusing, see An Intuitive Explanation of Bayesian Reasoning<sup>3</sup>.

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

Your strength as a rationalist<sup>4</sup> 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<sup>5</sup>. 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.

3. Page 28, 'An Intuitive Explanation of Bayes' Theorem'.

4. Page 126, 'Your Strength as a Rationalist'.

5. Page 108, 'Belief in Belief'.

### 3. Hindsight bias<sup>1</sup>

*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 group was told the outcome and also 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

---

<sup>1</sup>. [http://lesswrong.com/lw/il/hindsight\\_bias/](http://lesswrong.com/lw/il/hindsight_bias/)

*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<sup>2</sup>. Or as Fischhoff (1982) put it:

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.

---

<sup>2</sup>. Page 130, ‘Conservation of Expected Evidence’.

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<sup>3</sup>. *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/apos0075/05a00120/o>

## 4. Hindsight Devalues Science<sup>1</sup>

This excerpt<sup>2</sup> 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<sup>3</sup>. (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. (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<sup>4</sup>? Take a moment to think before continuing...

1. [http://lesswrong.com/lw/im/hindsight\\_devalues\\_science/](http://lesswrong.com/lw/im/hindsight_devalues_science/)

2. [http://csm1.som.ohio-state.edu/Music829C/hindsight\\_bias.html](http://csm1.som.ohio-state.edu/Music829C/hindsight_bias.html)

3. Page 133, 'Hindsight bias'.

4. Page 130, 'Conservation of Expected Evidence'.

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<sup>5</sup> 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 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<sup>6</sup> what you *really* would have expected.

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

---

5. Page 126, ‘Your Strength as a Rationalist’.

6. Page 130, ‘Conservation of Expected Evidence’.

## 5. Positive Bias: Look Into the Dark<sup>1</sup>

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”.

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.

---

<sup>1</sup>. [http://lesswrong.com/lw/iw/positive\\_bias\\_look\\_into\\_the\\_dark/](http://lesswrong.com/lw/iw/positive_bias_look_into_the_dark/)

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<sup>2</sup> 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 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*<sup>3</sup>—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

---

2. Page 147, ‘Fake Causality’.

3. Page 126, ‘Your Strength as a Rationalist’.

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<sup>4</sup>. 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?

---

4. Page 158, ‘The Futility of Emergence’.

## **Against Rationalization**



## I. Knowing About Biases Can Hurt People<sup>1</sup>

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<sup>2</sup> 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-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.

---

1. [http://lesswrong.com/lw/he/knowing\\_about\\_biases\\_can\\_hurt\\_people/](http://lesswrong.com/lw/he/knowing_about_biases_can_hurt_people/)

2. <http://www.sunysb.edu/polisci/mlodge/lodgemotivated.pdf>

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<sup>1</sup>

Politics is the mind-killer<sup>2</sup>. Debate is war, arguments are soldiers<sup>3</sup>. There is the temptation to search for ways to interpret every possible experimental result<sup>4</sup> 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.<sup>5</sup>

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<sup>6</sup>, 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 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.

---

1. [http://lesswrong.com/lw/ij/update\\_yourself\\_incrementally/](http://lesswrong.com/lw/ij/update_yourself_incrementally/)

2. Page 319, 'Politics is the Mind-Killer'.

3. Page 321, 'Policy Debates Should Not Appear One-Sided'.

4. Page 108, 'Belief in Belief'.

5. Page 130, 'Conservation of Expected Evidence'.

6. [http://lesswrong.com/lw/ig/i\\_defy\\_the\\_data/](http://lesswrong.com/lw/ig/i_defy_the_data/)

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<sup>7</sup>. If the theory is true, supporting evidence will 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<sup>8</sup>. 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

7. Page 130, ‘Conservation of Expected Evidence’.

8. [http://lesswrong.com/lw/ig/i\\_defy\\_the\\_data/](http://lesswrong.com/lw/ig/i_defy_the_data/)

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.<sup>9</sup> 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<sup>10</sup> 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 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<sup>11</sup> by keeping the beliefs you started with, after all.

---

9. Page 130, 'Conservation of Expected Evidence'.

10. [http://lesswrong.com/lw/i8/religions\\_claim\\_to\\_be\\_nondisprovable/](http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/)

11. [http://lesswrong.com/lw/b8/tsuyoku\\_naritai\\_i\\_want\\_to\\_become\\_stronger/](http://lesswrong.com/lw/b8/tsuyoku_naritai_i_want_to_become_stronger/)

### 3. One Argument Against An Army<sup>1</sup>

**Followup to:** Update Yourself Incrementally<sup>2</sup>

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<sup>3</sup>. 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 outweigh the first (by three arguments against one argument), so you keep your belief that Sylvania is responsible.

---

1. [http://lesswrong.com/lw/ik/one\\_argument\\_against\\_an\\_army/](http://lesswrong.com/lw/ik/one_argument_against_an_army/)

2. Page 508, ‘Update Yourself Incrementally’.

3. Page 508, ‘Update Yourself Incrementally’.

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,<sup>4</sup> 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 supporting 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. Page 508, ‘Update Yourself Incrementally’.

## 4. The Bottom Line<sup>1</sup>

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<sup>2</sup> of that physical ink with the physical boxes became fixed.

It may help to visualize a collection of worlds—Everett branches or Tegmark duplicates<sup>3</sup>—within which there is some 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 objec-

1. [http://lesswrong.com/lw/jt/the\\_bottom\\_line/](http://lesswrong.com/lw/jt/the_bottom_line/)

2. Page 86, ‘What is Evidence?’.

3. <http://arxiv.org/abs/astro-ph/0302131>

tive 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.

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?<sup>4</sup> for more on this topic.

---

4. Page 516, ‘What Evidence Filtered Evidence?’.

## 5. What Evidence Filtered Evidence?<sup>1</sup>

Yesterday I discussed the dilemma of the clever arguer<sup>2</sup>, 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<sup>3</sup> 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?

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

---

1. [http://lesswrong.com/lw/jt/what\\_evidence\\_filtered\\_evidence/](http://lesswrong.com/lw/jt/what_evidence_filtered_evidence/)

2. Page 513, ‘The Bottom Line’.

3. Page 89, ‘How Much Evidence Does It Take?’.

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<sup>4</sup> 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<sup>5</sup> screaming at each other<sup>6</sup>.

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<sup>7</sup> favorable signs and portents which you have not yet seen<sup>8</sup>. 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.

---

4. Page 324, 'The Scales of Justice, the Notebook of Rationality'.

5. Page 315, 'A Fable of Science and Politics'.

6. Page 319, 'Politics is the Mind-Killer'.

7. Page 511, 'One Argument Against An Army'.

8. Page 511, 'One Argument Against An Army'.

## 6. Rationalization<sup>1</sup>

**Followup to:** The Bottom Line<sup>2</sup>, What Evidence Filtered Evidence?<sup>3</sup>

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<sup>4</sup> 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 the bottom line, and then flowing backward to select<sup>5</sup> signs and portents for presentation<sup>6</sup> 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

---

1. <http://lesswrong.com/lw/ju/rationalization/>

2. Page 513, ‘The Bottom Line’.

3. Page 516, ‘What Evidence Filtered Evidence?’.

4. Page 513, ‘The Bottom Line’.

5. Page 516, ‘What Evidence Filtered Evidence?’.

6. Page 516, ‘What Evidence Filtered Evidence?’.

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<sup>7</sup>: 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. 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

---

7. Page 130, ‘Conservation of Expected Evidence’.

hypothesis, set out to *test* that hypothesis for the sake of *curiosity*—creating experiments that would drive their own beliefs in an unknown direction<sup>8</sup>.

If you genuinely don't know where you are going, you will probably feel quite curious about it. Curiosity is the first virtue<sup>9</sup>, 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 130, 'Conservation of Expected Evidence'.

9. <http://yudkowsky.net/virtues/>

## 7. A Rational Argument<sup>1</sup>

**Followup to:** The Bottom Line<sup>2</sup>, Rationalization<sup>3</sup>

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<sup>4</sup> under Bayes's Rule<sup>5</sup>?"

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 facts alone; they must condition on the additional fact<sup>6</sup> 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 can-

---

1. [http://lesswrong.com/lw/jw/a\\_rational\\_argument/](http://lesswrong.com/lw/jw/a_rational_argument/)

2. Page 513, 'The Bottom Line'.

3. Page 519, 'Rationalization'.

4. Page 86, 'What is Evidence?'.

5. Page 28, 'An Intuitive Explanation of Bayes' Theorem'.

6. Page 516, 'What Evidence Filtered Evidence?'.

dicate, 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.
- 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<sup>7</sup> 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 513, ‘The Bottom Line’.



## 8. Avoiding Your Belief's Real Weak Points<sup>1</sup>

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<sup>2</sup> 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<sup>3</sup>, I could have made a successful surprising prediction. Not long afterward, my 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<sup>4</sup> 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

---

1. [http://lesswrong.com/lw/jy/avoiding\\_your\\_beliefs\\_real\\_weak\\_points/](http://lesswrong.com/lw/jy/avoiding_your_beliefs_real_weak_points/)

2. Page 519, 'Rationalization'.

3. Page 126, 'Your Strength as a Rationalist'.

4. [http://en.wikipedia.org/wiki/Modern\\_Orthodox\\_Judaism](http://en.wikipedia.org/wiki/Modern_Orthodox_Judaism)

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<sup>5</sup>.

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 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

---

5. Page 636, ‘The Proper Use of Doubt’.

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<sup>6</sup> the standard defense will feel strengthening.

It probably feels really good, for example, to rehearse one's prescribed 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 First-born, 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 ti-

---

6. Page 511, 'One Argument Against An Army'.

tle 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<sup>7</sup>: 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 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

---

<sup>7</sup>. [http://lesswrong.com/lw/b8/tsuyoku\\_naritai\\_i\\_want\\_to\\_become\\_stronger/](http://lesswrong.com/lw/b8/tsuyoku_naritai_i_want_to_become_stronger/)

## 9. Motivated Stopping and Motivated Continuation<sup>1</sup>

**Followup to:** The Third Alternative<sup>2</sup>, The Meditation on Curiosity<sup>3</sup>

While I disagree with some views of the Fast and Frugal<sup>4</sup> 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 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

---

1. [http://lesswrong.com/lw/km/motivated\\_stopping\\_and\\_motivated\\_continuation/](http://lesswrong.com/lw/km/motivated_stopping_and_motivated_continuation/)

2. Page 589, ‘The Third Alternative’.

3. Page 640, ‘The Meditation on Curiosity’.

4. <http://fastandfrugal.com/>

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<sup>5</sup> 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<sup>6</sup>? 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.<sup>7</sup> (Apparent contradiction resolved as follows: Spending *one hour* discussing the problem with your mind carefully cleared of all conclusions, is different from waiting ten years on another \$20 million study.)

As for motivated stopping, it appears in every place a third alternative<sup>8</sup> is feared, and wherever you have an argument whose obvious counterargument<sup>9</sup> 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<sup>10</sup>, and so you’d rather not investigate how well your plan *really* worked, for fear of destroying the warm glow of moral satisfaction<sup>11</sup> you paid good money to purchase. It ap-

---

5. <http://bayes.cs.ucla.edu/BOOK-2K/>

6. [http://www.overcomingbias.com/2007/01/conspicuous\\_con.html](http://www.overcomingbias.com/2007/01/conspicuous_con.html)

7. Page 471, ‘Hold Off On Proposing Solutions’.

8. Page 589, ‘The Third Alternative’.

9. Page 640, ‘The Meditation on Curiosity’.

10. [http://lesswrong.com/lw/kb/cant\\_say\\_no\\_spending/](http://lesswrong.com/lw/kb/cant_say_no_spending/)

pears wherever your beliefs and anticipations get out of sync<sup>12</sup>, 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.

---

11. [http://lesswrong.com/lw/bw/scope\\_insensitivity/](http://lesswrong.com/lw/bw/scope_insensitivity/)

12. Page 108, 'Belief in Belief'.

## 10. A Case Study of Motivated Continuation<sup>1</sup>

I am not wholly unsympathetic to the many commenters in Torture vs. Dust Specks<sup>2</sup> who argued that it is preferable to inflict dust specks upon the eyes of  $3^{^\wedge}3$  (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 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.)

---

1. [http://lesswrong.com/lw/ko/a\\_case\\_study\\_of\\_motivated\\_continuation/](http://lesswrong.com/lw/ko/a_case_study_of_motivated_continuation/)

2. [http://lesswrong.com/lw/kn/torture\\_vs\\_dust\\_specks/](http://lesswrong.com/lw/kn/torture_vs_dust_specks/)

In “bioethics<sup>3</sup>” debates, you very often see experts on<sup>4</sup> 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<sup>5</sup> 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<sup>6</sup>?

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.”

Ducking uncomfortable choices is a dangerous habit of mind. There are certain times when it’s wise to suspend judgment<sup>7</sup> (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<sup>8</sup>.

Until you pick one interim best guess, the discomfort will consume your attention, distract you from the search, tempt you to

3. <http://www.singinst.org/blog/2007/10/21/should-ethicists-be-inside-or-outside-a-profession/>

4. [http://www.overcomingbias.com/2007/04/expert\\_at\\_versu.html](http://www.overcomingbias.com/2007/04/expert_at_versu.html)

5. [http://lesswrong.com/lw/ja/we\\_dont\\_really\\_want\\_your\\_participation/](http://lesswrong.com/lw/ja/we_dont_really_want_your_participation/)

6. Page 180, ‘Applause Lights’.

7. Page 471, ‘Hold Off On Proposing Solutions’.

8. Page 589, ‘The Third Alternative’.

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<sup>9</sup> 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*.

---

9. Page 529, 'Motivated Stopping and Motivated Continuation'.

## 11. Fake Justification<sup>1</sup>

Many Christians who've stopped really believing<sup>2</sup> now insist that they revere the Bible as a source of ethical advice. The standard atheist reply is given by Sam Harris<sup>3</sup>: "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<sup>4</sup>," 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<sup>5</sup>. But this is not really motivated stopping<sup>6</sup>. Calling it "motivated stopping" would imply that there was a search carried out in the first place.

In *The Bottom Line*<sup>7</sup>, 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

---

1. [http://lesswrong.com/lw/kq/fake\\_justification/](http://lesswrong.com/lw/kq/fake_justification/)

2. Page 108, 'Belief in Belief'.

3. [http://homepage.mac.com/pmcarlton/Harris\\_Sullivan\\_CompleteDebate.pdf](http://homepage.mac.com/pmcarlton/Harris_Sullivan_CompleteDebate.pdf)

4. <http://hardware.slashdot.org/article.pl?sid=07/03/26/197253>

5. Page 589, 'The Third Alternative'.

6. Page 520, 'Motivated Stopping and Motivated Continuation'.

7. Page 513, 'The Bottom Line'.

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<sup>8</sup> 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<sup>9</sup> of “I <heart> the Bible” is a historical misrepresentation of how the bottom line<sup>10</sup> really got there, like selling cat milk as cow milk. That is just not where the bottom line<sup>11</sup> 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<sup>12</sup>.

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

8. Page 640, ‘The Meditation on Curiosity’.

9. Page 513, ‘The Bottom Line’.

10. Page 513, ‘The Bottom Line’.

11. Page 513, ‘The Bottom Line’.

12. Page 469, ‘We Change Our Minds Less Often Than We Think’.

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. Fake Optimization Criteria<sup>1</sup>

**Followup to:** Fake Justification<sup>2</sup>, The Tragedy of Group Selectionism<sup>3</sup>

I've<sup>4</sup> previously<sup>5</sup> dwelt<sup>6</sup> in<sup>7</sup> considerable<sup>8</sup> length<sup>9</sup> upon<sup>10</sup> forms<sup>11</sup> of<sup>12</sup> rationalization<sup>13</sup> whereby<sup>14</sup> our<sup>15</sup> beliefs<sup>16</sup> appear<sup>17</sup> to<sup>18</sup> match<sup>19</sup> the<sup>20</sup> evidence<sup>21</sup> much<sup>22</sup> more<sup>23</sup> strongly<sup>24</sup> than<sup>25</sup> they<sup>26</sup> actually<sup>27</sup> do<sup>28</sup>. 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 principles<sup>29</sup> 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?

---

1. [http://lesswrong.com/lw/kz/fake\\_optimization\\_criteria/](http://lesswrong.com/lw/kz/fake_optimization_criteria/)

2. Page 535, 'Fake Justification'.

3. [http://lesswrong.com/lw/kw/the\\_tragedy\\_of\\_group\\_selectionism/](http://lesswrong.com/lw/kw/the_tragedy_of_group_selectionism/)

4. Page 133, 'Hindsight bias'.

5. Page 136, 'Hindsight Devalues Science'.

6. Page 126, 'Your Strength as a Rationalist'.

7. Page 119, 'Focus Your Uncertainty'.

8. Page 128, 'Absence of Evidence Is Evidence of Absence'.

9. Page 130, 'Conservation of Expected Evidence'.

10. Page 105, 'Making Beliefs Pay Rent (in Anticipated Experiences)'.

11. Page 108, 'Belief in Belief'.

12. Page 516, 'What Evidence Filtered Evidence?'.

13. Page 519, 'Rationalization'.

14. Page 114, 'Professing and Cheering'.

15. Page 117, 'Belief as Attire'.

16. [http://lesswrong.com/lw/i8/religions\\_claim\\_to\\_be\\_nondisprovable/](http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/)

17. Page 138, 'Fake Explanations'.

18. Page 140, 'Guessing the Teacher's Password'.

19. Page 147, 'Fake Causality'.

20. Page 151, 'Semantic Stopsigns'.

21. Page 86, 'What is Evidence?'.

22. Page 154, 'Mysterious Answers to Mysterious Questions'.

23. Page 158, 'The Futility of Emergence'.

24. Page 161, 'Say Not "Complexity"'.

25. Page 164, 'Positive Bias: Look Into the Dark'.

26. Page 505, 'Knowing About Biases Can Hurt People'.

27. Page 513, 'The Bottom Line'.

28. Page 516, 'What Evidence Filtered Evidence?'.

29. Page 535, 'Fake Justification'.

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<sup>30</sup> 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.)

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?

---

<sup>30</sup> Page 319, 'Politics is the Mind-Killer'.

Why, it's our old friend the alien god<sup>31</sup>, 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<sup>32</sup>. Natural selection doesn't *maximize* its criterion of inclusive genetic fitness—it's not that smart<sup>33</sup>. 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*<sup>34</sup>—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 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*<sup>35</sup>)

---

31. [http://lesswrong.com/lw/kr/an\\_alien\\_god/](http://lesswrong.com/lw/kr/an_alien_god/)

32. Page 315, 'A Fable of Science and Politics'.

33. [http://lesswrong.com/lw/kt/evolutions\\_are\\_stupid\\_but\\_work\\_anyway/](http://lesswrong.com/lw/kt/evolutions_are_stupid_but_work_anyway/)

34. [http://lesswrong.com/lw/kw/the\\_tragedy\\_of\\_group\\_selectionism/](http://lesswrong.com/lw/kw/the_tragedy_of_group_selectionism/)

35. [http://wiki.lesswrong.com/wiki/Evolution#Blog\\_posts\\_28sequence.29](http://wiki.lesswrong.com/wiki/Evolution#Blog_posts_28sequence.29)



## 13. Is That Your True Rejection?<sup>1</sup>

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<sup>2</sup> 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 launches 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<sup>3</sup>, at the *very first* moment the rejection occurred<sup>4</sup>.

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<sup>5</sup>?

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 every-

---

1. [http://lesswrong.com/lw/wj/is\\_that\\_your\\_true\\_rejection/](http://lesswrong.com/lw/wj/is_that_your_true_rejection/)

2. Page 144, ‘Science as Attire’.

3. Page 513, ‘The Bottom Line’.

4. Page 469, ‘We Change Our Minds Less Often Than We Think’.

5. Page 432, ‘Cultish Countercultishness’.

thing 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 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<sup>6</sup>: 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:

- Uncommon, but well-supported, scientific knowledge or math;
- Long inferential distances<sup>7</sup>;
- 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.

---

6. [http://www.overcomingbias.com/2006/12/agreeing\\_to\\_agr.html](http://www.overcomingbias.com/2006/12/agreeing_to_agr.html)

7. [http://lesswrong.com/lw/kg/expecting\\_short\\_inferential\\_distances/](http://lesswrong.com/lw/kg/expecting_short_inferential_distances/)

“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 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<sup>8</sup>. 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/](http://lesswrong.com/lw/nc/newcombs_problem_and_regret_of_rationality/)

## 14. Entangled Truths, Contagious Lies<sup>1</sup>

“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<sup>2</sup>, 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 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 con-

---

1. [http://lesswrong.com/lw/ur/entangled\\_truths\\_contagious\\_lies/](http://lesswrong.com/lw/ur/entangled_truths_contagious_lies/)

2. [http://lesswrong.com/lw/r8/and\\_the\\_winner\\_is\\_manyworlds/](http://lesswrong.com/lw/r8/and_the_winner_is_manyworlds/)

fused<sup>3</sup>" 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  $4.5 \times 10^{-31}$  m/s<sup>2</sup>, 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<sup>4</sup> 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<sup>5</sup>.

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<sup>6</sup> 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

3. Page 126, 'Your Strength as a Rationalist'.

4. <http://very.net/~nikolai/tb/coroner.html>

5. Page 465, 'How to Seem (and Be) Deep'.

6. Page 699, 'Reductionism'.

don't live in so righteous a universe<sup>7</sup>. 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<sup>8</sup>” 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 evolutionary biology can see the design signature of natural selection<sup>9</sup> 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.

---

7. [http://lesswrong.com/lw/uk/beyond\\_the\\_reach\\_of\\_god/](http://lesswrong.com/lw/uk/beyond_the_reach_of_god/)

8. [http://lesswrong.com/lw/m8/the\\_amazing\\_virgin\\_pregnancy/](http://lesswrong.com/lw/m8/the_amazing_virgin_pregnancy/)

9. [http://lesswrong.com/lw/kr/an\\_alien\\_god/](http://lesswrong.com/lw/kr/an_alien_god/)

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<sup>10</sup>. 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.

---

<sup>10</sup>. [http://lesswrong.com/lw/kj/no\\_one\\_knows\\_what\\_science\\_doesnt\\_know/](http://lesswrong.com/lw/kj/no_one_knows_what_science_doesnt_know/)

## 15. Of Lies and Black Swan Blowups<sup>1</sup>

**Followup to:** Entangled Truths, Contagious Lies<sup>2</sup>

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<sup>3</sup>.

That whole *suspiciously virtuous-sounding* theory<sup>4</sup> 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/](http://lesswrong.com/lw/9a/of_lies_and_black_swan_blowups/)

2. Page 546, 'Entangled Truths, Contagious Lies'.

3. [http://news.bbc.co.uk/2/hi/uk\\_news/magazine/7967982.stm](http://news.bbc.co.uk/2/hi/uk_news/magazine/7967982.stm)

4. Page 546, 'Entangled Truths, Contagious Lies'.

## 16. Dark Side Epistemology<sup>1</sup>

**Followup to:** Entangled Truths, Contagious Lies<sup>2</sup>

If you once tell a lie, the truth is ever after your enemy.

I have previously spoken<sup>3</sup> 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<sup>4</sup>” 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...

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.

---

1. [http://lesswrong.com/lw/uy/dark\\_side\\_epistemology/](http://lesswrong.com/lw/uy/dark_side_epistemology/)

2. Page 546, ‘Entangled Truths, Contagious Lies’.

3. Page 546, ‘Entangled Truths, Contagious Lies’.

4. [http://lesswrong.com/lw/kr/an\\_alien\\_god/](http://lesswrong.com/lw/kr/an_alien_god/)

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<sup>5</sup> 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<sup>6</sup>, which you must defeat. So you say: “I disagree! Not all 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<sup>7</sup>. ”

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<sup>8</sup> of the city, you have to walk through the streets<sup>9</sup> 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

5. [http://lesswrong.com/lw/rb/possibility\\_and\\_couldness/](http://lesswrong.com/lw/rb/possibility_and_couldness/)

6. Page 319, ‘Politics is the Mind-Killer’.

7. Page 108, ‘Belief in Belief’.

8. <http://yudkowsky.net/bayes/truth.html>

9. [http://lesswrong.com/lw/o5/the\\_second\\_law\\_of\\_thermodynamics\\_and\\_engines\\_of/](http://lesswrong.com/lw/o5/the_second_law_of_thermodynamics_and_engines_of/)

wrong. With extremely high probability<sup>10</sup>. 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?"<sup>11</sup> I don't have to believe if it's not absolutely certain<sup>12</sup>."

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 proverb 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<sup>13</sup>."

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<sup>14</sup>, 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<sup>15</sup> 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

10. [http://lesswrong.com/lw/06/perpetual\\_motion\\_beliefs/](http://lesswrong.com/lw/06/perpetual_motion_beliefs/)

11. Page 597, 'But There's Still A Chance, Right?'

12. Page 604, 'Absolute Authority'.

13. [http://lesswrong.com/lw/i8/religions\\_claim\\_to\\_be\\_nondisprovable/](http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/)

14. Page 108, 'Belief in Belief'.

15. Page 546, 'Entangled Truths, Contagious Lies'.

Art. Creating, in its stead, a vast complex of anti-law, rules of anti-thought, general justifications for believing the untrue.

Steven Kaas said<sup>16</sup>, “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.

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<sup>17</sup> of Deep Wisdom<sup>18</sup>—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

16. <http://www.acceleratingfuture.com/steven/?p=124>

17. Page 451, ‘Cached Thoughts’.

18. Page 465, ‘How to Seem (and Be) Deep’.

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<sup>19</sup>. The vast majority who go about repeating the Deep Wisdom are more duped than duplicitous, more self-deceived than deceiving. I think.

And it's surely not my intent to offer you a Fully General Counterargument<sup>20</sup>, 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<sup>21</sup>.

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?

---

19. Page 329, 'Are Your Enemies Innately Evil?'

20. Page 505, 'Knowing About Biases Can Hurt People'.

21. Page 485, 'The Genetic Fallacy'.

## 17. The Sacred Mundane<sup>1</sup>

### Followup to: Is Humanism a Religion-Substitute?<sup>2</sup>

So I was reading (around the first half of) Adam Frank's *The Constant Fire*, in preparation for my Bloggingheads dialogue<sup>3</sup> 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<sup>4</sup>—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 be somehow *different* from that of all the *other* people watching? Why, when we all have the same brain design<sup>5</sup>? 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*

1. [http://lesswrong.com/lw/57/the\\_sacred\\_mundane/](http://lesswrong.com/lw/57/the_sacred_mundane/)

2. Page 736, ‘Is Humanism A Religion-Substitute?’.

3. [http://lesswrong.com/lw/4i/bhtv\\_yudkowsky\\_adam\\_frank\\_on\\_religious\\_experience/](http://lesswrong.com/lw/4i/bhtv_yudkowsky_adam_frank_on_religious_experience/)

4. Page 739, ‘Scarcity’.

5. [http://lesswrong.com/lw/rl/the\\_psychological\\_unity\\_of\\_humankind/](http://lesswrong.com/lw/rl/the_psychological_unity_of_humankind/)

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<sup>6</sup>—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<sup>7</sup> that truly wishes to find answers. You will worship your own ignorance of the temporarily unanswered questions of your own generation—probably including<sup>8</sup> ones that are already answered<sup>9</sup>.

**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

---

6. Page 551, ‘Dark Side Epistemology’.

7. Page 640, ‘The Meditation on Curiosity’.

8. [http://lesswrong.com/lw/kj/no\\_one\\_knows\\_what\\_science\\_doesnt\\_know/](http://lesswrong.com/lw/kj/no_one_knows_what_science_doesnt_know/)

9. [http://lesswrong.com/lw/r5/the\\_quantum\\_physics\\_sequence/](http://lesswrong.com/lw/r5/the_quantum_physics_sequence/)

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<sup>10</sup> and demoted it to the “dull catalogue of mundane things” for the crime of its 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<sup>11</sup> 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.

---

<sup>10</sup>. Page 715, ‘Joy in the Merely Real’.

<sup>11</sup>. Page 715, ‘Joy in the Merely Real’.

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!<sup>12</sup>” 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 unfolding 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.

---

<sup>12</sup>. Page 628, ‘The Importance of Saying “Oops”’.



## **Against Doublethink**



## I. Singletthink<sup>1</sup>

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 simply 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

---

<sup>1</sup>. <http://lesswrong.com/lw/ko/singletthink/>

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,<sup>2</sup> 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 565, ‘Doublethink (Choosing to be Biased)’.

## **2. Doublethink (Choosing to be Biased)<sup>1</sup>**

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.'

'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.

---

<sup>1.</sup> [http://lesswrong.com/lw/jc/doublethink\\_choosing\\_to\\_be\\_biased/](http://lesswrong.com/lw/jc/doublethink_choosing_to_be_biased/)

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<sup>2</sup>, *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<sup>3</sup> it—though I have just made that more difficult for you by pointing out the difference. (You’re 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.

---

2. Page 356, ‘Human Evil and Muddled Thinking’.

3. Page 108, ‘Belief in Belief’.

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<sup>4</sup>.

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 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<sup>5</sup>. 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.

---

4. [http://en.wikipedia.org/wiki/Neglect\\_of\\_probability](http://en.wikipedia.org/wiki/Neglect_of_probability)

5. <http://lesswrong.com/lw/j5/availability/>

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.

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<sup>1</sup>

**Followup to:** Belief in Belief<sup>2</sup>

<sup>3</sup>

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<sup>4</sup> 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 *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

---

1. [http://lesswrong.com/lw/r/no\\_really\\_ive\\_deceived\\_myself/](http://lesswrong.com/lw/r/no_really_ive_deceived_myself/)

2. Page 108, ‘Belief in Belief’.

3. Page 108, ‘Belief in Belief’.

4. [http://lesswrong.com/lw/yp/pretending\\_to\\_be\\_wise/](http://lesswrong.com/lw/yp/pretending_to_be_wise/)

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<sup>1</sup>

**Continuation of:** No, Really, I've Deceived Myself<sup>2</sup>

**Followup to:** Dark Side Epistemology<sup>3</sup>

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<sup>4</sup>—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 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:

---

1. [http://lesswrong.com/lw/s/belief\\_in\\_selfdeception/](http://lesswrong.com/lw/s/belief_in_selfdeception/)

2. Page 569, ‘No, Really, I've Deceived Myself’.

3. Page 108, ‘Belief in Belief’.

4. Page 640, ‘The Meditation on Curiosity’.

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<sup>5</sup>,” but I’m still having trouble coming to grips<sup>6</sup> 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<sup>7</sup> and the Lady 3rd Kiritsugu<sup>8</sup>, 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<sup>9</sup>, the inside of your mind ends up as an Escher painting.

5. Page 99, ‘The Lens That Sees Its Flaws’.

6. [http://lesswrong.com/lw/hb/think\\_like\\_reality/](http://lesswrong.com/lw/hb/think_like_reality/)

7. [http://lesswrong.com/lw/ys/the\\_babyeating.aliens\\_18/](http://lesswrong.com/lw/ys/the_babyeating.aliens_18/)

8. [http://lesswrong.com/lw/y7/the\\_super\\_happy\\_people\\_38/](http://lesswrong.com/lw/y7/the_super_happy_people_38/)

9. Page 551, ‘Dark Side Epistemology’.

(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<sup>10</sup>.)

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.

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<sup>11</sup> that she had built around beliefs such as “I benefit from believing people are nicer than they actually are.”

---

<sup>10</sup>. Page 465, ‘How to Seem (and Be) Deep’.

<sup>11</sup>. Page 551, ‘Dark Side Epistemology’.

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<sup>12</sup> 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 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<sup>13</sup>, 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<sup>14</sup> and its proponents still haven't abandoned it.

---

12. Page 565, 'Doublethink (Choosing to be Biased)'.

13. Page 108, 'Belief in Belief'.

14. Page 772, 'Zombies! Zombies?'.

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.

## 5. Moore's Paradox<sup>1</sup>

**Followup to:** Belief in Self-Deception<sup>2</sup>

Moore's Paradox<sup>3</sup> is the standard term for saying “It’s raining outside but I don’t believe that it is.” HT to painquale on MetaFilter.<sup>4</sup>

I think I understand Moore’s Paradox a bit better now, after reading some of the comments on Less Wrong. Jimrandomh<sup>5</sup> 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<sup>6</sup> that causes confused thinking (or maybe it just reflects pre-existing confusion).

So: in the original example<sup>7</sup>, “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 quot-

---

1. [http://lesswrong.com/lw/tf/moores\\_paradox/](http://lesswrong.com/lw/tf/moores_paradox/)

2. Page 571, ‘Belief in Self-Deception’.

3. [http://en.wikipedia.org/wiki/Moore%027s\\_paradox](http://en.wikipedia.org/wiki/Moore%027s_paradox)

4. [#2477702](http://www.metafilter.com/79752/More-Right-was-too-political)

5. [http://lesswrong.com/lw/r/no\\_really\\_iive\\_deceived\\_myself/#ga](http://lesswrong.com/lw/r/no_really_iive_deceived_myself/#ga)

6. Page 253, ‘Fallacies of Compression’.

7. Page 571, ‘Belief in Self-Deception’.

ed 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<sup>8</sup>; 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<sup>9</sup>, and providing another mechanism whereby people can be simultaneously right and wrong<sup>10</sup>.

Likewise Kurige<sup>11</sup> who wrote:

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?

---

8. Page 108, 'Belief in Belief'.

9. Page 571, 'Belief in Self-Deception'.

10. [http://lesswrong.com/lw/rd/simultaneously\\_right\\_and\\_wrong/](http://lesswrong.com/lw/rd/simultaneously_right_and_wrong/)

11. [http://lesswrong.com/lw/r/no\\_really\\_ive\\_deceived\\_myself/#gk](http://lesswrong.com/lw/r/no_really_ive_deceived_myself/#gk)

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<sup>1</sup>

**Followup to:** Moore's Paradox<sup>2</sup>, Doublethink

<sup>3</sup>

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<sup>4</sup>:

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<sup>5</sup>...

...how can you still *believe* it?" I helplessly want to say.

Or<sup>6</sup>:

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"?

Well... apparently not.

One part of this puzzle may be my explanation of Moore's Paradox<sup>7</sup> ("It's raining, but I don't believe it is")—that people introspectively mistake positive affect attached to a quoted belief, for actual credulity.

---

1. [http://lesswrong.com/lw/1o/dont\\_believe\\_youll\\_selfdeceive/](http://lesswrong.com/lw/1o/dont_believe_youll_selfdeceive/)

2. Page 576, 'Moore's Paradox'.

3. Page 576, 'Moore's Paradox'.

4. [http://lesswrong.com/lw/r/no\\_really\\_ive\\_deceived\\_myself/#gk](http://lesswrong.com/lw/r/no_really_ive_deceived_myself/#gk)

5. Page 565, 'Doublethink (Choosing to be Biased)'.

6. [http://lesswrong.com/lw/1f/moores\\_paradox/#u3](http://lesswrong.com/lw/1f/moores_paradox/#u3)

7. Page 576, 'Moore's Paradox'.

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<sup>8</sup>, and even then, it may take more work to get the implications on a *gut level*.

I realize now that when I wrote<sup>9</sup> “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<sup>10</sup>, belief in belief<sup>11</sup>, or belief in self-deception<sup>12</sup>.

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—

---

8. Page 86, ‘What is Evidence?’.

9. Page 576, ‘Moore’s Paradox’.

10. Page 576, ‘Moore’s Paradox’.

11. Page 108, ‘Belief in Belief’.

12. Page 571, ‘Belief in Self-Deception’.

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.



## **Overly Convenient Excuses**



## I. The Proper Use of Humility<sup>1</sup>

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 spacesuit, 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.

---

<sup>1</sup>. [http://lesswrong.com/lw/gq/the\\_proper\\_use\\_of\\_humility/](http://lesswrong.com/lw/gq/the_proper_use_of_humility/)

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<sup>2</sup> 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!

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 gen-

---

2. [http://www.sunysb.edu/polsci/mlodge/taber\\_lodge\\_ajps.pdf](http://www.sunysb.edu/polsci/mlodge/taber_lodge_ajps.pdf)

uflected 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<sup>1</sup>

*“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<sup>2</sup>, 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<sup>3</sup>, the fallacy of the excluded middle, or the package-deal fallacy<sup>4</sup>. 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 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

---

1. [http://lesswrong.com/lw/bu/the\\_third\\_alternative/](http://lesswrong.com/lw/bu/the_third_alternative/)

2. <http://yudkowsky.net/virtues/>

3. [http://en.wikipedia.org/wiki/False\\_dilemma](http://en.wikipedia.org/wiki/False_dilemma)

4. [http://en.wikipedia.org/wiki/Package-deal\\_fallacy](http://en.wikipedia.org/wiki/Package-deal_fallacy)

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<sup>5</sup>, 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 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

---

5. Page 319, 'Politics is the Mind-Killer'.

Liars who calculate an optimal new Noble Lie; they keep whatever lie they started with. Better stop that search fast!

To do better<sup>6</sup>, 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<sup>7</sup> 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?

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.

---

6. [http://lesswrong.com/lw/b8/tsuyoku\\_naritai\\_i\\_want\\_to\\_become\\_stronger/](http://lesswrong.com/lw/b8/tsuyoku_naritai_i_want_to_become_stronger/)

7. Page 585, ‘The Proper Use of Humility’.

### 3. Privileging the Hypothesis<sup>1</sup>

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<sup>2</sup> which is not legal evidence<sup>3</sup> 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<sup>4</sup>. It’s human nature to look for confirmation rather than disconfirmation<sup>5</sup>. 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-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<sup>6</sup>”, which is that if you have a billion boxes only one of

1. [http://lesswrong.com/lw/19m/privileging\\_the\\_hypothesis/](http://lesswrong.com/lw/19m/privileging_the_hypothesis/)

2. [http://wiki.lesswrong.com/wiki/Rational\\_evidence](http://wiki.lesswrong.com/wiki/Rational_evidence)

3. [http://wiki.lesswrong.com/wiki/Rational\\_evidence](http://wiki.lesswrong.com/wiki/Rational_evidence)

4. [http://wiki.lesswrong.com/wiki/Locate\\_the\\_hypothesis](http://wiki.lesswrong.com/wiki/Locate_the_hypothesis)

5. Page 164, ‘Positive Bias: Look Into the Dark’.

6. [http://wiki.lesswrong.com/wiki/Locate\\_the\\_hypothesis](http://wiki.lesswrong.com/wiki/Locate_the_hypothesis)

which contains a diamond (the truth), and your detectors only provide 1 bit of evidence<sup>7</sup> 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<sup>8</sup> rather than legal evidence<sup>9</sup>, 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<sup>10</sup>, 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<sup>11</sup> or single-world quantum mechanics<sup>12</sup>. But, said Scott, we might encounter *future* evidence in favor of single-world quantum mechanics, and many-worlds still has the open question of the Born probabilities<sup>13</sup>.

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, discontinuous, 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<sup>14</sup>. Something that unphysical is not worth *saying out loud*

7. [http://wiki.lesswrong.com/wiki/Amount\\_of\\_evidence](http://wiki.lesswrong.com/wiki/Amount_of_evidence)

8. [http://wiki.lesswrong.com/wiki/Rational\\_evidence](http://wiki.lesswrong.com/wiki/Rational_evidence)

9. [http://wiki.lesswrong.com/wiki/Rational\\_evidence](http://wiki.lesswrong.com/wiki/Rational_evidence)

10. <http://www.bloggingheads.tv/diavlogs/21857>

11. [http://lesswrong.com/lw/q6/collapse\\_postulates/](http://lesswrong.com/lw/q6/collapse_postulates/)

12. [http://lesswrong.com/lw/q8/many\\_worlds\\_one\\_best\\_guess/](http://lesswrong.com/lw/q8/many_worlds_one_best_guess/)

13. [http://lesswrong.com/lw/py/the\\_born\\_probabilities/](http://lesswrong.com/lw/py/the_born_probabilities/)

14. [http://lesswrong.com/lw/q6/collapse\\_postulates/](http://lesswrong.com/lw/q6/collapse_postulates/)

or even *thinking about as a possibility* without a rather large weight of evidence<sup>15</sup>—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 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<sup>16</sup> 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”.

---

15. [http://wiki.lesswrong.com/wiki/Amount\\_of\\_evidence](http://wiki.lesswrong.com/wiki/Amount_of_evidence)

16. <http://wiki.lesswrong.com/wiki/Anti-epistemology>

Just as Occam's Razor<sup>17</sup> 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<sup>18</sup> 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<sup>19</sup>, 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. 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

---

17. [http://wiki.lesswrong.com/wiki/Occam%027s\\_razor](http://wiki.lesswrong.com/wiki/Occam%027s_razor)

18. [http://wiki.lesswrong.com/wiki/Burdensome\\_details](http://wiki.lesswrong.com/wiki/Burdensome_details)

19. [http://lesswrong.com/lw/o6/perpetual\\_motion\\_beliefs/](http://lesswrong.com/lw/o6/perpetual_motion_beliefs/)

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<sup>20</sup>.

---

<sup>20</sup>. [http://wiki.lesswrong.com/wiki/Privilege\\_the\\_hypothesis](http://wiki.lesswrong.com/wiki/Privilege_the_hypothesis)

## 4. But There's Still A Chance, Right?<sup>1</sup>

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 chance, but worth keeping track of." You can see this in the OB lottery debate<sup>2</sup>, where someone said, "There's a big difference between zero chance of winning and epsilon chance of winning," and

---

1. [http://lesswrong.com/lw/ml/but\\_theres\\_still\\_a\\_chance\\_right/](http://lesswrong.com/lw/ml/but_theres_still_a_chance_right/)

2. [http://lesswrong.com/lw/bm/new\\_improved\\_lottery/](http://lesswrong.com/lw/bm/new_improved_lottery/)

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<sup>3</sup> 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 ignore 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 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?

---

3. Page 335, ‘Reversed Stupidity Is Not Intelligence’.

## 5. The Fallacy of Gray<sup>1</sup>

**Followup to:** Tsuyoku Naritai<sup>2</sup>, But There's Still A Chance Right?<sup>3</sup>

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 million 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:

“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,

---

1. [http://lesswrong.com/lw/mm/the\\_fallacy\\_of\\_gray/](http://lesswrong.com/lw/mm/the_fallacy_of_gray/)

2. [http://lesswrong.com/lw/b8/tsuyoku\\_naritai\\_i\\_want\\_to\\_become\\_stronger/](http://lesswrong.com/lw/b8/tsuyoku_naritai_i_want_to_become_stronger/)

3. Page 597, ‘But There’s Still A Chance, Right?’.

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 everyday life and matters like personal self-improvement. That which I cannot *eliminate* may be well worth *reducing*.

Or consider this exchange between Robin Hanson<sup>4</sup> and Tyler Cowen<sup>5</sup>. 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 the-

---

4. <http://www.overcomingbias.com/2007/12/economist-judgm.html>

5. <http://www.marginalrevolution.com/marginalrevolution/2007/12/how-my-views-di.html>

ory, 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<sup>6</sup>; but for those who strive to do better<sup>7</sup>, 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.")

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<sup>8</sup> to heal your daughter of cancer, that too is not a certainty. But they are rather different

---

6. Page 180, 'Applause Lights'.

7. [http://lesswrong.com/lw/b8/tsuyoku\\_naritai\\_i\\_want\\_to\\_become\\_stronger/](http://lesswrong.com/lw/b8/tsuyoku_naritai_i_want_to_become_stronger/)

8. Page 108, 'Belief in Belief'.

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<sup>9</sup>.

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-triumphant 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, 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.

---

9. Page 122, ‘The Virtue of Narrowness’.

**Addendum:** G<sup>10</sup> points us to Asimov's The Relativity of Wrong<sup>11</sup>: "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/bbo](http://lesswrong.com/lw/mm/the_fallacy_of_gray/bbo)

11. <http://chem.tufts.edu/AnswersInScience/RelativityofWrong.htm>

## 6. Absolute Authority<sup>1</sup>

**Followup to:** But There's Still A Chance Right?<sup>2</sup>, The Fallacy of Gray<sup>3</sup>

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<sup>4</sup>. 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 counterarguments<sup>5</sup> 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.

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

---

1. [http://lesswrong.com/lw/mn/absolute\\_authority/](http://lesswrong.com/lw/mn/absolute_authority/)

2. Page 597, ‘But There’s Still A Chance, Right?’.

3. Page 599, ‘The Fallacy of Gray’.

4. [http://lesswrong.com/lw/kg/expecting\\_short\\_inferential\\_distances/](http://lesswrong.com/lw/kg/expecting_short_inferential_distances/)

5. Page 321, ‘Policy Debates Should Not Appear One-Sided’.

any flaw in them would destroy their authority utterly; so any trace of doubt is a sin, and claiming certainty<sup>6</sup> 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<sup>7</sup> 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*.)

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

---

6. Page 114, 'Professing and Cheering'.

7. Page 119, 'Focus Your Uncertainty'.

this knowledge that the unenlightened ones set above ours, and why? It's probably some musty old scroll that has been contradicted eighteen 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<sup>8</sup> 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<sup>9</sup>, "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 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."

---

8. [http://lesswrong.com/lw/kg/expecting\\_short\\_inferential\\_distances/](http://lesswrong.com/lw/kg/expecting_short_inferential_distances/)

9. [http://lesswrong.com/lw/kg/expecting\\_short\\_inferential\\_distances/](http://lesswrong.com/lw/kg/expecting_short_inferential_distances/)

- “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<sup>10</sup> 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<sup>11</sup> must be correct for 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.<sup>12</sup>

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.

---

<sup>10</sup>. Page 335, 'Reversed Stupidity Is Not Intelligence'.

<sup>11</sup>. [http://lesswrong.com/lw/jk/burdensome\\_details/](http://lesswrong.com/lw/jk/burdensome_details/)

<sup>12</sup>. <http://www.nizkor.org/features/fallacies/appeal-to-consequences.html>

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 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.<sup>13</sup>”) 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 INFINITE-LY certain” rather than “We are not certain”. For the latter case,

---

<sup>13</sup>. [http://lesswrong.com/lw/bm/new\\_improved\\_lottery/](http://lesswrong.com/lw/bm/new_improved_lottery/)

in ordinary discourse, suggests you know some specific reason for doubt.

## 7. How to Convince Me That $2 + 2 = 3$ <sup>1</sup>

In “What is Evidence?”, I wrote<sup>2</sup>:

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<sup>3</sup>:

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 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  $XXX - XX = XX$ .

---

1. [http://lesswrong.com/lw/jr/bow\\_to\\_convince\\_me\\_that\\_2\\_2\\_3/](http://lesswrong.com/lw/jr/bow_to_convince_me_that_2_2_3/)

2. Page 86, ‘What is Evidence?’.

3. [http://lesswrong.com/lw/jl/what\\_is\\_evidence/f7b](http://lesswrong.com/lw/jl/what_is_evidence/f7b)

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<sup>4</sup>.

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.

There’s really only two possibilities, for a belief of fact<sup>5</sup>—either the belief got there via a mind-reality entangling process<sup>6</sup>, or not. If not, the belief can’t be correct except by coincidence. For beliefs with the slightest shred of internal complexity<sup>7</sup> (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.

---

4. Page 126, ‘Your Strength as a Rationalist’.

5. Page 625, ‘Feeling Rational’.

6. Page 86, ‘What is Evidence?’.

7. Page 95, ‘Occam’s Razor’.

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?

## 8. Infinite Certainty<sup>1</sup>

**Followup to:** How To Convince Me That  $2 + 2 = 3^2$ , Absolute Authority<sup>3</sup>

In Absolute Authority<sup>4</sup>, 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 map and the territory<sup>6</sup>. Given the seeming absolute stability and universality of physical laws<sup>7</sup>, 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<sup>8</sup>, but whether you have any knowledge of this 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

---

1. [http://lesswrong.com/lw/mo/infinite\\_certainty/](http://lesswrong.com/lw/mo/infinite_certainty/)

2. Page 92, 'How to Convince Me That  $2 + 2 = 3$ '.

3. Page 604, 'Absolute Authority'.

4. Page 604, 'Absolute Authority'.

5. Page 92, 'How to Convince Me That  $2 + 2 = 3$ '.

6. <http://yudkowsky.net/bayes/truth.html>

7. [http://lesswrong.com/lw/br/universal\\_law/](http://lesswrong.com/lw/br/universal_law/)

8. <http://ansuz.sooke.bc.ca/bonobo-conspiracy/?i=369>

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,  $\text{SSo} + \text{SSo} = \text{SSSSo}$ " 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  $2 + 2 = 3$ <sup>9</sup>, 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 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<sup>10</sup>—because I've seen that when I blurt this out to people without proper preparation, they thereafter use it as a Fully General Counterargument<sup>11</sup>, which somehow leaps to mind whenever they have to

9. Page 92, 'How to Convince Me That  $2 + 2 = 3$ '.

10. <http://singinst.org/Biases.pdf>

11. Page 505, 'Knowing About Biases Can Hurt People'.

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.

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 99.99999% 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 ( $1 - 1/\text{googolplex}$ ) 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  $3^{\wedge\wedge} 3^{12}$ .

But even a confidence of ( $1 - 1/3^{\wedge\wedge} 3$ ) 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 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."

---

12. [http://lesswrong.com/lw/kn/torture\\_vs\\_dust\\_specks/](http://lesswrong.com/lw/kn/torture_vs_dust_specks/)

## 9. 0 And 1 Are Not Probabilities<sup>1</sup>

### Followup to: Infinite Certainty<sup>2</sup> <sup>3</sup>

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, " $5 + \text{infinity} = \text{infinity}$ ", 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 " $\text{infinity} - \text{infinity} = 5$ ". 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.

(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.)

---

1. [http://lesswrong.com/lw/mp/o\\_and\\_1\\_are\\_not\\_probabilities/](http://lesswrong.com/lw/mp/o_and_1_are_not_probabilities/)

2. Page 614, 'Infinite Certainty'.

3. Page 604, 'Absolute Authority'.

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  $2/5 = 0.40$ ). Then I convert back into a probability, if I like, and get  $(0.4 / 1.4) = 2/7 \approx 29\%$ .

So odds are more manageable for Bayesian updates—if you use probabilities, you've got to deploy Bayes's Theorem<sup>4</sup> in its 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

---

4. Page 28, ‘An Intuitive Explanation of Bayes' Theorem’.

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  $-40 \text{ db} + 20 \text{ db} = -20 \text{ db}$ , 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<sup>5</sup> requires infinitely strong evidence, just as infinite absurdity requires infinitely strong counterevidence.

---

5. Page 614, 'Infinite Certainty'.

Furthermore, all sorts of standard theorems in probability have special cases if you try to plug  $1$ s or  $0$ s 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<sup>6</sup> of coming up some number between  $1$  and  $6$ . The die might land on its edge; or get struck by 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<sup>7</sup>. (I told him not to do it again.)

---

<sup>6</sup>. Page 614, ‘Infinite Certainty’.

<sup>7</sup>. <http://www.spaceandgames.com/?p=27>



**Letting Go**



## I. Feeling Rational<sup>1</sup>

A popular belief about “rationality” is that rationality opposes all emotion—that all our sadness and all our joy are automatically antilogical 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 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

---

<sup>1</sup>. [http://lesswrong.com/lw/hp/feeling\\_rational/](http://lesswrong.com/lw/hp/feeling_rational/)

goblin tied your shoelaces. The criterion of rationality spreads virtually, 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<sup>2</sup>, 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

---

2. <http://yudkowsky.net/yebuda.html>

confess strong emotions, and you're probably uncomfortable hearing them. But I know, now, that it is rational to feel.

## 2. The Importance of Saying “Oops”<sup>1</sup>

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 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.

---

<sup>1</sup>. [http://lesswrong.com/lw/ig/the\\_importance\\_of\\_saying\\_oops/](http://lesswrong.com/lw/ig/the_importance_of_saying_oops/)

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<sup>2</sup>. 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 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.

---

2. [http://lesswrong.com/lw/b8/tsuyoku\\_naritai\\_i\\_want\\_to\\_become\\_stronger/](http://lesswrong.com/lw/b8/tsuyoku_naritai_i_want_to_become_stronger/)

Better to swallow the entire bitter pill in one terrible gulp.

### 3. The Crackpot Offer<sup>1</sup>

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.)

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

---

<sup>1</sup>. [http://lesswrong.com/lw/j8/the\\_crackpot\\_offer/](http://lesswrong.com/lw/j8/the_crackpot_offer/)

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<sup>2</sup>, 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 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 be-

---

2. Page 628, ‘The Importance of Saying “Oops”’.

ings make mistakes; it happens, that's all. Say "oops<sup>3</sup>", and get on with your life.

---

3. Page 628, 'The Importance of Saying "Oops".

## 4. Just Lose Hope Already<sup>1</sup>

Casey Serin, a 24-year-old web programmer with no prior experience in real estate, owes banks 2.2 million dollars<sup>2</sup> 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<sup>3</sup> 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<sup>4</sup>.

While LTCM raked in giant profits over its first three years, in 1998 the inefficiencies 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.

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*

---

1. [http://lesstrong.com/lw/gx/just\\_lose\\_hope\\_already/](http://lesstrong.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](http://en.wikipedia.org/wiki/Long-Term_Capital_Management)

*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<sup>1</sup>

Once, when I was holding forth upon the Way<sup>2</sup>, 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: *Ab, 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<sup>3</sup>; there is no curiosity that does not *want* an answer. But if you obtain an answer, 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

---

1. [http://lesswrong.com/lw/ib/the\\_proper\\_use\\_of\\_doubt/](http://lesswrong.com/lw/ib/the_proper_use_of_doubt/)

2. <http://yudkowsky.net/virtues/>

3. <http://yudkowsky.net/virtues/>

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<sup>4</sup> 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<sup>5</sup> a religion (not just believe in<sup>6</sup> it), then why would you tell your novices to consider doubts that must 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<sup>7</sup>. Making a great public display of doubt

---

4. Page 117, ‘Belief as Attire’.

5. Page 105, ‘Making Beliefs Pay Rent (in Anticipated Experiences)’.

6. Page 108, ‘Belief in Belief’.

7. Page 585, ‘The Proper Use of Humility’.

to convince yourself<sup>8</sup> that you are a rationalist, will do around as much good as wearing a lab coat.

To avoid professing<sup>9</sup> 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.

---

8. Page 108, 'Belief in Belief'.

9. Page 114, 'Professing and Cheering'.

## 6. You Can Face Reality<sup>1</sup>

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.)

---

<sup>1</sup>. [http://lesswrong.com/lw/id/you\\_can\\_face\\_reality/](http://lesswrong.com/lw/id/you_can_face_reality/)

## 7. The Meditation on Curiosity<sup>1</sup>

“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<sup>2</sup> 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<sup>3</sup>—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 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.

---

1. [http://lesswrong.com/lw/jz/the\\_meditation\\_on\\_curiosity/](http://lesswrong.com/lw/jz/the_meditation_on_curiosity/)

2. Page 589, ‘The Third Alternative’.

3. [http://lesswrong.com/lw/bw/scope\\_insensitivity/](http://lesswrong.com/lw/bw/scope_insensitivity/)

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<sup>4</sup>. 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,<sup>5</sup> must be dared as a danger if we want to become stronger<sup>6</sup>. 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.”

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,<sup>7</sup> that you are flinching away from—it’s not all negative thinking.

4. Page 315, ‘A Fable of Science and Politics’.

5. Page 505, ‘Knowing About Biases Can Hurt People’.

6. [http://lesswrong.com/lw/b8/tsuyoku\\_naritai\\_i\\_want\\_to\\_become\\_stronger/](http://lesswrong.com/lw/b8/tsuyoku_naritai_i_want_to_become_stronger/)

7. Page 525, ‘Avoiding Your Belief’s Real Weak Points’.

It should also help to meditate on Conservation of Expected Evidence<sup>8</sup>. 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<sup>9</sup>, rather than its weak points? Are you rehearsing the evidence<sup>10</sup>?

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,*

---

8. Page 130, ‘Conservation of Expected Evidence’.

9. Page 525, ‘Avoiding Your Belief’s Real Weak Points’.

10. Page 511, ‘One Argument Against An Army’.

*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<sup>ii</sup>.

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.

---

ii. Page 639, 'You Can Face Reality'.

## 8. Something to Protect<sup>1</sup>

**Followup to:** Tsuyoku Naritai<sup>2</sup>, Circular Altruism<sup>3</sup>

In the gestalt of (ahem) Japanese<sup>4</sup> fiction<sup>5</sup>, 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<sup>6</sup>—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 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.

---

1. [http://lesswrong.com/lw/nb/something\\_to\\_protect/](http://lesswrong.com/lw/nb/something_to_protect/)

2. [http://lesswrong.com/lw/b8/tsuyoku\\_naritai\\_i\\_want\\_to\\_become\\_stronger/](http://lesswrong.com/lw/b8/tsuyoku_naritai_i_want_to_become_stronger/)

3. [http://lesswrong.com/lw/n3/circular\\_altruism/](http://lesswrong.com/lw/n3/circular_altruism/)

4. [http://lesswrong.com/lw/m7/zen\\_and\\_the\\_art\\_of\\_rationality/](http://lesswrong.com/lw/m7/zen_and_the_art_of_rationality/)

5. Page 459, ‘The Logical Fallacy of Generalization from Fictional Evidence’.

6. <http://tv tropes.org/pmwiki/pmwiki.php/Main/KilledOffForReal>

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<sup>7</sup> 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.

---

7. Page 83, ‘Why truth? And...’.

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<sup>8</sup>.

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 contradiction, 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<sup>9</sup>.

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<sup>10</sup>.

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.

---

8. Page 83, ‘Why truth? And...’.

9. Page 565, ‘Doublethink (Choosing to be Biased)’.

10. [http://lesswrong.com/lw/ng/the\\_intuitions\\_behind\\_utilitarianism/](http://lesswrong.com/lw/ng/the_intuitions_behind_utilitarianism/)

In Circular Altruism<sup>11</sup>, 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<sup>12</sup>.

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.

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!”

---

11. [http://lesswrong.com/lw/n3/circular\\_altruism/](http://lesswrong.com/lw/n3/circular_altruism/)

12. [http://lesswrong.com/lw/n9/the\\_intuitions\\_behind\\_utilitarianism/](http://lesswrong.com/lw/n9/the_intuitions_behind_utilitarianism/)

(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<sup>13</sup> 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 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:

---

<sup>13</sup>. Page 428, 'Lonely Dissent'.

“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<sup>14</sup> 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<sup>15</sup>. Learn how to multiply<sup>16</sup>, 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<sup>17</sup> for its own sake.

---

14. Page 406, ‘Every Cause Wants To Be A Cult’.

15. Page 465, ‘How to Seem (and Be) Deep’.

16. [http://lesswrong.com/lw/n9/the\\_intuitions\\_behind\\_utilitarianism/](http://lesswrong.com/lw/n9/the_intuitions_behind_utilitarianism/)

17. [http://lesswrong.com/lw/mt/beautiful\\_probability/](http://lesswrong.com/lw/mt/beautiful_probability/)

## 9. No One Can Exempt You From Rationality's Laws<sup>1</sup>

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<sup>2</sup> into a map that reflects the territory<sup>3</sup>. The principles of rationality are laws<sup>4</sup> in the same sense as the second law of thermodynamics: obtaining a reliable belief requires a calculable amount of entangled evidence<sup>5</sup>, just as reliably cooling the contents of a refrigerator requires a calculable minimum of free energy.

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<sup>6</sup> street

1. [http://lesswrong.com/lw/k1/no\\_one\\_can\\_exempt\\_you\\_from\\_rationalitys\\_laws/](http://lesswrong.com/lw/k1/no_one_can_exempt_you_from_rationalitys_laws/)

2. Page 86, ‘What is Evidence?’.

3. <http://sl4.org/wiki/TheSimpleTruth>

4. [http://lesswrong.com/lw/br/universal\\_law/](http://lesswrong.com/lw/br/universal_law/)

5. Page 89, ‘How Much Evidence Does It Take?’.

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<sup>7</sup>, 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 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

---

6. [http://lesswrong.com/lw/jk/burdensome\\_details/](http://lesswrong.com/lw/jk/burdensome_details/)

7. [http://lesswrong.com/lw/b9/tsuyoku\\_vs\\_the\\_egalitarian\\_instinct/](http://lesswrong.com/lw/b9/tsuyoku_vs_the_egalitarian_instinct/)

things, at least not in the wiser countries. But no one can revoke the law<sup>8</sup> that you need evidence<sup>9</sup> to generate *accurate* beliefs<sup>10</sup>. 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"<sup>11</sup> 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/](http://lesswrong.com/lw/hr/universal_law/)

9. Page 86, 'What is Evidence?'.

10. <http://sl4.org/wiki/TheSimpleTruth>

11. Page 519, 'Rationalization'.

## 10. Leave a Line of Retreat<sup>1</sup>

“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. 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.”

---

<sup>1</sup>. [http://lesswrong.com/lw/o4/leave\\_a\\_line\\_of\\_retreat/](http://lesswrong.com/lw/o4/leave_a_line_of_retreat/)

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<sup>2</sup> and the Litany of Tarski<sup>3</sup>. 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.

---

2. Page 525, 'Avoiding Your Belief's Real Weak Points'.

3. Page 640, 'The Meditation on Curiosity'.

(*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.)

## II. Crisis of Faith<sup>1</sup>

**Followup to:** Make an Extraordinary Effort<sup>2</sup>, The Meditation on Curiosity<sup>3</sup>, Avoiding Your Belief's Real Weak Points<sup>4</sup>

“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<sup>5</sup> mind to relinquish, if the skepticism of a ten-year-old were applied without evasion<sup>6</sup>. 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<sup>7</sup> 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<sup>8</sup>. We change our minds less often than we think.<sup>9</sup>

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.

---

1. [http://lesswrong.com/lw/ur/crisis\\_of\\_faith/](http://lesswrong.com/lw/ur/crisis_of_faith/)

2. [http://lesswrong.com/lw/u0/make\\_an\\_extraordinary\\_effort/](http://lesswrong.com/lw/u0/make_an_extraordinary_effort/)

3. Page 640, ‘The Meditation on Curiosity’.

4. Page 525, ‘Avoiding Your Belief’s Real Weak Points’.

5. Page 384, ‘Affective Death Spirals’.

6. Page 525, ‘Avoiding Your Belief’s Real Weak Points’.

7. Page 95, ‘Occam’s Razor’.

8. Page 636, ‘The Proper Use of Doubt’.

9. Page 469, ‘We Change Our Minds Less Often Than We Think’.

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<sup>10</sup>, 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<sup>11</sup>, where you just cough up points *you already thought of long before*, is *exactly* the style of thinking that keeps people within their current religions. If you stay with your cached thoughts<sup>12</sup>, if your brain fills in the obvious answer so fast that you can’t see originally<sup>13</sup>, you surely will not be able to conduct a crisis of faith.

Maybe it’s just a question of not enough people reading “Godel, Escher, Bach” at a sufficiently young age, but I’ve noticed that a large fraction of the population—even technical folk—have trouble following<sup>14</sup> arguments that go this meta. On my more pessimistic days I wonder if the camel has two humps.

---

<sup>10</sup>. Page 329, ‘Are Your Enemies Innately Evil?’.

<sup>11</sup>. Page 511, ‘One Argument Against An Army’.

<sup>12</sup>. Page 451, ‘Cached Thoughts’.

<sup>13</sup>. Page 457, ‘Original Seeing’.

<sup>14</sup>. [http://lesswrong.com/lw/b6/chronophone\\_motivations/](http://lesswrong.com/lw/b6/chronophone_motivations/)

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.

"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 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<sup>15</sup> or the Simulation Hypothesis<sup>16</sup>. 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

---

15. [http://lesswrong.com/lw/q3/decoherence\\_is\\_simple/](http://lesswrong.com/lw/q3/decoherence_is_simple/)

16. <http://www.simulation-argument.com/simulation.html>

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<sup>17</sup> 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.

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<sup>18</sup> 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:

---

17. [http://lesswrong.com/lw/u0/make\\_an\\_extraordinary\\_effort/](http://lesswrong.com/lw/u0/make_an_extraordinary_effort/)

18. Page 485, ‘The Genetic Fallacy’.

- A belief has long remained in your mind;
- It is surrounded by a cloud of known arguments and refutations;
- You have sunk costs<sup>19</sup> in it (time, money, public declarations);
- The belief has emotional consequences<sup>20</sup> (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<sup>21</sup> for all sorts of dangers, 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<sup>22</sup>. When a retina sends the same signal regardless of the photons entering it, we call that eye blind<sup>23</sup>.

### 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

19. [http://en.wikipedia.org/wiki/Sunk\\_cost](http://en.wikipedia.org/wiki/Sunk_cost)

20. Page 625, 'Feeling Rational'.

21. Page 432, 'Cultish Countercultishness'.

22. Page 513, 'The Bottom Line'.

23. Page 86, 'What is Evidence?'.

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 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<sup>24</sup>—One of the first temptations in a crisis of faith is to doubt the strongest points of your belief, so that you can rehearse<sup>25</sup> 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<sup>26</sup>—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<sup>27</sup> may help you Update Yourself Incrementally<sup>28</sup>: For every *single* point that you consider, and each element of new argument and new evidence, you should not expect your beliefs to shift

---

24. Page 525, ‘Avoiding Your Belief’s Real Weak Points’.

25. Page 511, ‘One Argument Against An Army’.

26. Page 640, ‘The Meditation on Curiosity’.

27. Page 130, ‘Conservation of Expected Evidence’.

28. Page 508, ‘Update Yourself Incrementally’.

more (on average) in one direction than another—thus you can be truly curious each time about how it will go.

- Cached Thoughts<sup>29</sup> and Pirsig's Original Seeing<sup>30</sup>, to prevent standard thoughts from rushing in and completing the pattern.
- The Litany of Gendlin<sup>31</sup> and the Litany of Tarski<sup>32</sup>: 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<sup>33</sup>, 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<sup>34</sup>: 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 written by human hands, but still clings to the idea that it contains indispensable ethical wisdom<sup>35</sup>.)

---

29. Page 451, 'Cached Thoughts'.

30. Page 457, 'Original Seeing'.

31. Page 639, 'You Can Face Reality'.

32. Page 640, 'The Meditation on Curiosity'.

33. [http://lesswrong.com/lw/u0/make\\_an\\_extraordinary\\_effort/](http://lesswrong.com/lw/u0/make_an_extraordinary_effort/)

34. Page 485, 'The Genetic Fallacy'.

35. [http://lesswrong.com/lw/i8/religions\\_claim\\_to\\_be\\_nondisprovable/](http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/)

- The Importance of Saying “Oops”<sup>36</sup>—it really is less painful to swallow the entire bitter pill in one terrible gulp.
- Singlethink<sup>37</sup>, the opposite of doublethink. See the thoughts you flinch away from, that appear in the corner of your mind for just a moment <sup>38</sup>before you refuse to think them. If you become aware of what you are not thinking, you can think it.
- Affective Death Spirals<sup>39</sup> and Resist the Happy Death Spiral<sup>40</sup>. 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<sup>41</sup>.
- Hold Off On Proposing Solutions<sup>42</sup> 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<sup>43</sup>; and try for five minutes before giving up<sup>44</sup>, both generally, and especially when pursuing the devil’s point of view.

And these standard techniques are particularly relevant:

- The sequence on The Bottom Line<sup>45</sup> and Rationalization<sup>46</sup>, which explains why it is always wrong to selectively argue one side of a debate.
- Positive Bias<sup>47</sup> and motivated skepticism<sup>48</sup> and motivated stopping<sup>49</sup>, lest you selectively look for support,

36. Page 628, ‘The Importance of Saying “Oops”’.

37. Page 563, ‘Singlethink’.

38. Page 563, ‘Singlethink’.

39. Page 384, ‘Affective Death Spirals’.

40. Page 386, ‘Resist the Happy Death Spiral’.

41. Page 321, ‘Policy Debates Should Not Appear One-Sided’.

42. Page 471, ‘Hold Off On Proposing Solutions’.

43. Page 469, ‘We Change Our Minds Less Often Than We Think’.

44. [http://lesswrong.com/lw/ui/use\\_the\\_try\\_harder\\_luke/](http://lesswrong.com/lw/ui/use_the_try_harder_luke/)

45. Page 513, ‘The Bottom Line’.

46. Page 519, ‘Rationalization’.

47. Page 164, ‘Positive Bias: Look Into the Dark’.

48. Page 505, ‘Knowing About Biases Can Hurt People’.

selectively look for counter-counterarguments, and selectively stop the argument before it gets dangerous. Missing alternatives<sup>50</sup> are a special case of stopping. A special case of motivated skepticism is fake humility<sup>51</sup> where you bashfully confess that no one can know<sup>52</sup> something you would rather not know. Don't selectively demand too much authority<sup>53</sup> of counterarguments.

- Beware of Semantic Stopsigns<sup>54</sup>, Applause Lights<sup>55</sup>, and the choice to Explain/Worship/Ignore<sup>56</sup>.
- Feel the weight of Burdensome Details<sup>57</sup>; 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 *ishoukenmei*—the lifelong uncompromising effort to be so incredibly rational that you rise above the level of stupid damn mistakes. It's when you 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!

---

49. Page 529, 'Motivated Stopping and Motivated Continuation'.

50. Page 589, 'The Third Alternative'.

51. Page 585, 'The Proper Use of Humility'.

52. [http://lesswrong.com/lw/kj/no\\_one\\_knows\\_what\\_science\\_doesnt\\_know/](http://lesswrong.com/lw/kj/no_one_knows_what_science_doesnt_know/)

53. Page 604, 'Absolute Authority'.

54. Page 151, 'Semantic Stopsigns'.

55. Page 180, 'Applause Lights'.

56. Page 175, 'Explain/Worship/Ignore?'.

57. [http://lesswrong.com/lw/jk/burdensome\\_details/](http://lesswrong.com/lw/jk/burdensome_details/)

## 12. The Ritual<sup>1</sup>

**Followup to:** The Failures of Eld Science<sup>2</sup>, Crisis of Faith<sup>3</sup>

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.”

“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—

---

1. [http://lesswrong.com/lw/us/the\\_ritual/](http://lesswrong.com/lw/us/the_ritual/)

2. [http://lesswrong.com/lw/q9/the\\_failures\\_of\\_eld\\_science/](http://lesswrong.com/lw/q9/the_failures_of_eld_science/)

3. Page 657, ‘Crisis of Faith’.

"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 Jeffreysai 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.

Jeffreysai 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, Jeffreysai 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.



## **Part V**

### **Reductionism**

*How to take reality apart into pieces... and live in that universe, where we have always lived, without feeling disappointed about the fact that complicated things are made of simpler things.*



## I. Dissolving the Question<sup>1</sup>

**Followup to:** How an Algorithm Feels From the Inside<sup>2</sup>, Feel the Meaning<sup>3</sup>, Replace the Symbol with the Substance<sup>4</sup>

“If a tree falls in the forest<sup>5</sup>, but no one hears it, does it make a sound?”

I didn’t *answer* that question. I didn’t pick a position, “Yes!” or “No!”, and defend it. Instead I went off and deconstructed<sup>6</sup> the human algorithm for processing words, even going so far as to sketch an illustration<sup>7</sup> of a neural network. At the end, I hope, there was no question left—not even the feeling of a question.

Many philosophers—particularly amateur philosophers, and ancient philosophers—share a dangerous instinct: If you give them a question, they try to answer it.

Like, say, “Do we have free will?”

The dangerous instinct of philosophy is to marshal the arguments in favor, and marshal the arguments against, and weigh them up, and publish them in a prestigious journal of philosophy, and so finally conclude: “Yes, we must have free will,” or “No, we cannot possibly have free will.”

Some philosophers are wise enough to recall the warning that most philosophical disputes are really disputes over the meaning of a word, or confusions generated by using different meanings for the same word in different places<sup>8</sup>. So they try to define very precisely what they mean by “free will”, and then ask again, “Do we have free will? Yes or no?”

A philosopher wiser yet, may suspect that the confusion about “free will” shows the notion itself is flawed. So they pursue the Traditional Rationalist course: They argue that “free will” is inherently self-contradictory, or meaningless because it has no testable conse-

---

1. [http://lesswrong.com/lw/of/dissolving\\_the\\_question/](http://lesswrong.com/lw/of/dissolving_the_question/)

2. Page 223, ‘How An Algorithm Feels From Inside’.

3. Page 234, ‘Feel the Meaning’.

4. Page 248, ‘Replace the Symbol with the Substance’.

5. Page 228, ‘Disputing Definitions’.

6. Page 223, ‘How An Algorithm Feels From Inside’.

7. Page 217, ‘Neural Categories’.

8. Page 298, ‘Variable Question Fallacies’.

quences<sup>9</sup>. And then they publish these devastating observations in a prestigious philosophy journal.

But *proving that* you are confused may not make you feel any *less* confused. Proving that a question is meaningless may not help you any more than answering it.

The philosopher's instinct is to find the most defensible position, publish it, and move on. But the "naive" view, the instinctive view, is a fact about human psychology. You can prove that free will is impossible until the Sun goes cold, but this leaves an unexplained fact of cognitive science: If free will doesn't exist, what goes on inside the head of a human being who thinks it does? This is not a rhetorical question!

It is a fact about human psychology that people think they have free will. Finding a more defensible *philosophical position* doesn't change, or explain, that *psychological fact*. Philosophy may lead you to *reject* the concept, but rejecting a concept is not the same as understanding the cognitive algorithms behind it.

You could look at the Standard Dispute<sup>10</sup> over "If a tree falls in the forest, and no one hears it, does it make a sound?", and you could do the Traditional Rationalist thing: Observe that the two don't disagree on any point of anticipated experience<sup>11</sup>, and triumphantly declare the argument pointless. That happens to be correct in this particular case; but, as *a question of cognitive science*, why did the arguers make that mistake in the first place?

The key idea of the heuristics and biases program is that the *mistakes* we make, often reveal far more about our underlying cognitive algorithms than our correct answers. So (I asked myself, once upon a time) what kind of mind design<sup>12</sup> corresponds to the mistake of arguing<sup>13</sup> about trees falling in deserted forests?

The cognitive algorithms we use, *are* the way the world feels<sup>14</sup>. And these cognitive algorithms may not have a one-to-one correspondence with reality—not even macroscopic reality, to say nothing

9. Page 105, 'Making Beliefs Pay Rent (in Anticipated Experiences)'.

10. Page 228, 'Disputing Definitions'.

11. Page 105, 'Making Beliefs Pay Rent (in Anticipated Experiences)'.

12. Page 234, 'Feel the Meaning'.

13. Page 228, 'Disputing Definitions'.

14. Page 223, 'How An Algorithm Feels From Inside'.

ing of the true quarks. There can be things in the mind that cut skew to the world.

For example, there can be a dangling unit<sup>15</sup> in the center of a neural network<sup>16</sup>, which does not correspond to any real thing, or any real property of any real thing, existent anywhere in the real world. This dangling unit is often useful as a shortcut in computation<sup>17</sup>, which is why we have them. (Metaphorically speaking. Human neurobiology is surely far more complex<sup>18</sup>.)

This dangling unit *feels like* an unresolved question, even after every answerable query<sup>19</sup> is answered. No matter how much anyone proves to you that no difference of anticipated experience depends on the question, you're left wondering: "But does the falling tree *really* make a sound, or not?"

But once you understand *in detail* how your brain generates the *feeling* of the question—once you realize that your feeling of an unanswered question, corresponds to an illusory central unit wanting to know whether it should fire, even after all the edge units are clamped at known values—or better yet, you understand the technical workings of Naive Bayes<sup>20</sup>—then you're done. Then there's no lingering feeling of confusion, no vague sense of dissatisfaction.

If there is *any* lingering feeling of a remaining unanswered question, or of having been fast-talked into something, then this is a sign that you have not dissolved the question. A vague dissatisfaction<sup>21</sup> should be as much warning as a shout. *Really* dissolving the question doesn't leave anything behind.

A triumphant thundering refutation of free will, an absolutely unarguable proof that free will cannot exist, feels very *satisfying*—a grand cheer<sup>22</sup> for the home team<sup>23</sup>. And so you may not notice that—as a point of cognitive science—you do not have a full and satisfactory descriptive explanation of how each intuitive sensation arises, point by point.

---

15. Page 223, 'How An Algorithm Feels From Inside'.

16. Page 217, 'Neural Categories'.

17. Page 288, 'Conditional Independence, and Naive Bayes'.

18. Page 288, 'Conditional Independence, and Naive Bayes'.

19. Page 213, 'Disguised Queries'.

20. Page 288, 'Conditional Independence, and Naive Bayes'.

21. Page 126, 'Your Strength as a Rationalist'.

22. Page 114, 'Professing and Cheering'.

23. [http://lesswrong.com/lw/mg/the\\_twoparty\\_swindle/](http://lesswrong.com/lw/mg/the_twoparty_swindle/)

You may not even want to admit your ignorance, of this point of cognitive science, because that would feel like a score against Your Team. In the midst of smashing all foolish beliefs of free will, it would seem like a concession to the opposing side to concede that you've left anything unexplained.

And so, perhaps, you'll come up with a just-so evolutionary-psychological<sup>24</sup> argument that hunter-gatherers who believed in free will, were more likely to take a positive outlook on life, and so outreproduce other hunter-gatherers—to give one example of a completely bogus explanation. If you say this, you are *arguing that* the brain generates an illusion of free will—but you are not *explaining how*. You are trying to dismiss the opposition by deconstructing its motives—but in the story you tell, the illusion of free will is a brute fact. You have not taken the illusion apart to see the wheels and gears.

Imagine that in the Standard Dispute about a tree falling in a deserted forest, you first prove that no difference of anticipation exists, and then go on to hypothesize, “But perhaps people who said that arguments were meaningless were viewed as having conceded, and so lost social status, so now we have an instinct to argue about the meanings of words.” That’s *arguing that* or *explaining why* a confusion exists. Now look at the neural network structure in Feel the Meaning<sup>25</sup>. That’s *explaining how*, disassembling the confusion into smaller pieces which are not themselves confusing. See the difference?

Coming up with good hypotheses about cognitive algorithms (or even hypotheses that hold together for half a second) is a good deal harder than just refuting a philosophical confusion. Indeed, it is an entirely different art. Bear this in mind, and you should feel less embarrassed to say, “I know that what you say can’t possibly be true, and I can prove it. But I cannot write out a flowchart which shows how your brain makes the mistake, so I’m not done yet, and will continue investigating.”

I say all this, because it sometimes seems to me that at least 20% of the real-world effectiveness of a skilled rationalist comes from not stopping too early<sup>26</sup>. If you keep asking questions, you’ll

---

<sup>24</sup>. [http://lesswrong.com/lw/mk/a\\_failed\\_justso\\_story/](http://lesswrong.com/lw/mk/a_failed_justso_story/)

<sup>25</sup>. Page 234, ‘Feel the Meaning’.

<sup>26</sup>. Page 640, ‘The Meditation on Curiosity’.

get to your destination eventually. If you decide too early that you've found an answer, you won't.

The challenge, above all, is to notice when you are confused—even if it just feels like a little tiny bit of confusion—and even if there's someone standing across from you, *insisting* that humans have free will, and *smirking* at you, and the fact that you don't know *exactly* how the cognitive algorithms work, has *nothing to do* with the searing folly of their position...

But when you can lay out the cognitive algorithm in sufficient detail that you can walk through the thought process, step by step, and describe how each intuitive perception arises—decompose the confusion into smaller pieces not themselves confusing—*then* you're done.

So be warned that you may *believe* you're done, when all you have is a mere triumphant refutation of a mistake<sup>27</sup>.

But when you're *really* done, you'll *know* you're done.<sup>28</sup> Dissolving the question is an unmistakable feeling—once you experience it, and, having experienced it, resolve not to be fooled again. Those who dream do not know they dream, but when you wake you know you are awake.<sup>29</sup>

Which is to say: When you're done, you'll know you're done, but unfortunately the reverse implication does not hold.

So here's your homework problem: What kind of cognitive algorithm, as felt from the inside, would generate the observed debate about “free will”?

Your assignment is not to argue about whether people have free will, or not.

Your assignment is not to argue that free will is compatible with determinism, or not.

Your assignment is not to argue that the question is ill-posed, or that the concept is self-contradictory, or that it has no testable consequences.

You are not asked to invent an evolutionary explanation of how people who believed in free will would have reproduced; nor an account of how the concept of free will seems suspiciously congruent

---

27. Page 335, ‘Reversed Stupidity Is Not Intelligence’.

28. [http://lesswrong.com/lw/gr/the\\_modesty\\_argument/](http://lesswrong.com/lw/gr/the_modesty_argument/)

29. [http://lesswrong.com/lw/gr/the\\_modesty\\_argument/](http://lesswrong.com/lw/gr/the_modesty_argument/)

with bias X. Such are mere attempts to *explain why* people believe in “free will”, not *explain how*.

Your homework assignment is to write a stack trace of the internal algorithms of the human mind as they produce the intuitions that power the whole damn philosophical argument.

This is one of the first real challenges I tried as an aspiring rationalist, once upon a time. One of the easier conundrums, relatively speaking. May it serve you likewise.

## 2. Wrong Questions<sup>1</sup>

**Followup to:** Dissolving the Question<sup>2</sup>, Mysterious Answers to Mysterious Questions<sup>3</sup>

Where the mind cuts against reality's grain, it generates *wrong questions*—questions that cannot possibly be answered *on their own terms*, but only dissolved<sup>4</sup> by understanding the cognitive algorithm that generates the *perception* of a question.

One good cue that you're dealing with a “wrong question” is when you cannot even *imagine* any concrete, specific state of how-the-world-is that would answer the question. When it doesn't even seem *possible* to answer the question.

Take the Standard Definitional Dispute<sup>5</sup>, for example, about the tree falling in a deserted forest. Is there any way-the-world-could-be—any state of affairs—that corresponds to the word “sound” *really meaning* only acoustic vibrations, or *really meaning* only auditory experiences?

(“Why, yes,” says the one, “it is the state of affairs where ‘sound’ means acoustic vibrations.” So Taboo<sup>6</sup> the word ‘means’, and ‘represents’, and all similar synonyms, and describe again: How can the world be, what state of affairs, would make one side right, and the other side wrong?)

Or if that seems too easy, take free will: What concrete state of affairs, whether in deterministic physics, or in physics with a dice-rolling random component, could ever correspond to having free will?

And if *that* seems too easy, then ask “Why does anything exist at all?”, and then tell me what a satisfactory answer to that question would even *look like*.

And no, I don't know the answer to that last one. But I *can* guess one thing, based on my previous experience with unanswerable questions. The answer will not consist of some grand triumphant First Cause. The question will go away as a result of some

---

1. [http://lesswrong.com/lw/og/wrong\\_questions/](http://lesswrong.com/lw/og/wrong_questions/)

2. Page 673, ‘Dissolving the Question’.

3. Page 154, ‘Mysterious Answers to Mysterious Questions’.

4. Page 673, ‘Dissolving the Question’.

5. Page 228, ‘Disputing Definitions’.

6. Page 248, ‘Replace the Symbol with the Substance’.

insight into how my mental algorithms run skew to reality, after which I will understand how the question itself was wrong from the beginning—how the question itself assumed the fallacy, contained the skew.

Mystery exists in the mind, not in reality. If I am ignorant about a phenomenon, that is a fact about my state of mind, not a fact about the phenomenon itself. All the more so, if it seems like no possible answer can exist: Confusion exists in the map, not in the territory. *Unanswerable* questions do not mark places where magic enters the universe. They mark places where your mind runs skew to reality.

Such questions *must* be dissolved. Bad things happen when you try to answer them. It inevitably generates the worst sort of Mysterious Answer to a Mysterious Question<sup>7</sup>: The one where you come up with seemingly strong arguments for your Mysterious Answer, but the “answer” doesn’t let you make any new predictions even in retrospect, and the phenomenon still possesses the same sacred inexplicability that it had at the start.

I could guess, for example, that the answer to the puzzle of the First Cause is that nothing *does* exist—that the whole concept of “existence” is bogus. But if you sincerely believed that, would you be any less confused? Me neither.

But the wonderful thing about *unanswerable* questions is that they are *always* solvable, at least in my experience. What went through Queen Elizabeth I’s mind, first thing in the morning, as she woke up on her fortieth birthday? As I can easily *imagine* answers to this question, I can readily see that I may never be able to *actually* answer it, the true information having been lost in time.

On the other hand, “Why does anything exist at all?” seems *so* absolutely impossible that I can infer that I am just confused, one way or another, and the truth probably isn’t all that complicated in an absolute sense, and once the confusion goes away I’ll be able to see it.

This may seem counterintuitive if you’ve never solved an unanswerable question, but I assure you that it *is* how these things work.

Coming tomorrow: A simple trick for handling “wrong questions”.

---

7. Page 154, ‘Mysterious Answers to Mysterious Questions’.



### 3. Righting a Wrong Question<sup>1</sup>

**Followup to:** How an Algorithm Feels from the Inside<sup>2</sup>, Dissolving the Question<sup>3</sup>, Wrong Questions<sup>4</sup>

When you are faced with an *unanswerable* question—a question to which it seems impossible to even *imagine* an answer—there is a simple trick which can turn the question solvable.

Compare:

- “Why do I have free will?”
- “Why do I think I have free will?”

The nice thing about the second question is that it is *guaranteed* to have a real answer, *whether or not* there is any such thing as free will. Asking “Why do I have free will?” or “Do I have free will?” sends you off thinking about tiny details of the laws of physics, so distant from the macroscopic level that you couldn’t begin to see them with the naked eye. And you’re asking “Why is X the case?” where X may not be *coherent*, let alone the case.

“Why do I *think* I have free will?”, in contrast, is guaranteed answerable. You do, in fact, believe you have free will. This belief seems far more solid and graspable than the ephemerality of free will. And there is, *in fact*, some nice solid chain of cognitive cause and effect leading up to this belief.

If you’ve already outgrown free will, choose one of these substitutes:

- “Why does time move forward instead of backward?” versus “Why do I think time moves forward instead of backward?”
- “Why was I born as myself rather than someone else?” versus “Why do I think I was born as myself rather than someone else?”
- “Why am I conscious?” versus “Why do I think I’m conscious?”
- “Why does reality exist?” versus “Why do I think reality exists?”

---

1. [http://lesswrong.com/lw/ob/righting\\_a\\_wrong\\_question/](http://lesswrong.com/lw/ob/righting_a_wrong_question/)

2. Page 223, ‘How An Algorithm Feels From Inside’.

3. Page 673, ‘Dissolving the Question’.

4. Page 679, ‘Wrong Questions’.

The beauty of this method is that it works *whether or not* the question is confused. As I type this, I am wearing socks. I could ask “Why am I wearing socks?” or “Why do I believe I’m wearing socks?” Let’s say I ask the second question. Tracing back the chain of causality, I find:

- I believe I’m wearing socks, because I can see socks on my feet.
- I see socks on my feet, because my retina is sending sock signals to my visual cortex.
- My retina is sending sock signals, because sock-shaped light is impinging on my retina.
- Sock-shaped light impinges on my retina, because it reflects from the socks I’m wearing.
- It reflects from the socks I’m wearing, because I’m wearing socks.
- I’m wearing socks because I put them on.
- I put socks on because I believed that otherwise my feet would get cold.
- &c.

Tracing back the chain of causality, step by step, I discover that my belief that I’m wearing socks is fully explained by the fact that I’m wearing socks. This is right and proper, as you cannot gain information about something without interacting with it<sup>5</sup>.

On the other hand, if I see a mirage of a lake in a desert, the correct causal explanation of my vision does not involve the fact of any actual lake in the desert. In this case, my belief in the lake is not just *explained*, but *explained away*.

But *either way*, the belief itself is a real phenomenon taking place in the real universe—psychological events are events—and its causal history can be traced back.

“Why is there a lake in the middle of the desert?” may fail if there is no lake to be explained. But “Why do I *perceive* a lake in the middle of the desert?” always has a causal explanation, one way or the other.

Perhaps someone will see an opportunity to be clever, and say: “Okay. I believe in free will because I have free will. There, I’m done.” Of course it’s not that easy.

---

<sup>5</sup>. [http://lesswrong.com/lw/o6/perpetual\\_motion\\_beliefs/](http://lesswrong.com/lw/o6/perpetual_motion_beliefs/)

My perception of socks on my feet, is an event in the visual cortex. The workings of the visual cortex can be investigated by cognitive science, should they be confusing.

My retina receiving light is not a mystical sensing procedure, a magical sock detector that lights in the presence of socks for no explicable reason; there are mechanisms that can be understood in terms of biology. The photons entering the retina can be understood in terms of optics. The shoe's surface reflectance can be understood in terms of electromagnetism and chemistry. My feet getting cold can be understood in terms of thermodynamics.

So it's not as easy as saying, "I believe I have free will because I have it—there, I'm done!" You have to be able to break the causal chain into smaller steps, and explain the steps in terms of elements not themselves confusing.

The mechanical interaction of my retina with my socks is quite clear, and can be described in terms of non-confusing components like photons and electrons. Where's the free-will-sensor in your brain, and how does it detect the presence or absence of free will? How does the sensor interact with the sensed event, and what are the mechanical details of the interaction?

If your belief does derive from valid observation of a real phenomenon, we will eventually reach that fact, if we start tracing the causal chain backward from your belief.

If what you are really seeing is your own confusion, tracing back the chain of causality will find an algorithm that runs skew to reality<sup>6</sup>.

Either way, the question is guaranteed to have an answer. You even have a nice, concrete place to begin tracing—your belief, sitting there solidly in your mind.

Cognitive science may not seem so lofty and glorious as metaphysics. But at least questions of cognitive science are *solvab*le. Finding an answer may not be *easy*, but at least an answer *exists*.

Oh, and also: the idea that cognitive science is not so lofty and glorious as metaphysics is simply wrong. Some readers are beginning to notice this, I hope.

---

6. Page 673, 'Dissolving the Question'.

## 4. Mind Projection Fallacy<sup>1</sup>

**Followup to:** How an Algorithm Feels From Inside<sup>2</sup>

Monsterwithgirl\_2

In the dawn days of science fiction, alien invaders would occasionally kidnap a girl in a torn dress and carry her off for intended ravishing, as lovingly depicted on many ancient magazine covers. Oddly enough, the aliens never go after men in torn shirts.

Would a non-humanoid alien, with a different evolutionary history and evolutionary psychology<sup>3</sup>, sexually desire a human female? It seems rather unlikely. To put it mildly.

People don't make mistakes like that by deliberately reasoning: "All possible minds are likely to be wired pretty much the same way, therefore a bug-eyed monster will find human females attractive." Probably the artist did not even think to ask whether an alien *perceives* human females as attractive. Instead, a human female in a torn dress *is sexy*—inherently so, as an intrinsic property.

They who went astray did not think about the alien's evolutionary history; they focused on the woman's torn dress. If the dress were not torn, the woman would be less sexy; the alien monster doesn't enter into it.

Apparently we instinctively represent Sexiness as a direct attribute of the Woman object, `Woman.sexiness`, like `Woman.height` or `Woman.weight`.

If your brain uses that data structure, or something metaphorically similar to it, then from the inside<sup>4</sup> it feels like sexiness is an

---

1. [http://lesswrong.com/lw/oi/mind\\_projection\\_fallacy/](http://lesswrong.com/lw/oi/mind_projection_fallacy/)

2. Page 223, 'How An Algorithm Feels From Inside'.

3. [http://lesswrong.com/lw/l1/evolutionary\\_psychology/](http://lesswrong.com/lw/l1/evolutionary_psychology/)

4. Page 223, 'How An Algorithm Feels From Inside'.

inherent property of the woman, not a property of the alien looking at the woman. Since the woman *is attractive*, the alien monster will be *attracted* to her—isn’t that logical?

E. T. Jaynes used the term Mind Projection Fallacy<sup>5</sup> to denote the error of projecting your own mind’s properties into the external world. Jaynes, as a late grand master of the Bayesian Conspiracy, was most concerned with the mistreatment of *probabilities* as inherent properties of objects, rather than states of partial knowledge in some particular mind. More about this shortly.

But the Mind Projection Fallacy generalizes as an error. It is in the argument over the real meaning of the word *sound*<sup>6</sup>, and in the magazine cover of the monster carrying off a woman in the torn dress, and Kant’s declaration that space by its very nature is flat, and Hume’s definition of *a priori*<sup>7</sup> ideas as those “discoverable by the mere operation of thought, without dependence on what is anywhere existent in the universe”...

(Incidentally, I once read an SF story about a human male who entered into a sexual relationship with a sentient alien plant of appropriately squishy fronds; discovered that it was an androecious<sup>8</sup> (male) plant; agonized about this for a bit; and finally decided that it didn’t really matter at that point. And in Foglio and Pollotta’s *Illegal Aliens*, the humans land on a planet inhabited by sentient insects, and see a movie advertisement showing a human carrying off a bug in a delicate chiffon dress. Just thought I’d mention that.)

---

5. <http://citeseer.ist.psu.edu/6330.html>

6. Page 228, ‘Disputing Definitions’.

7. [http://lesswrong.com/lw/k2/a\\_priori/](http://lesswrong.com/lw/k2/a_priori/)

8. [http://en.wikipedia.org/wiki/Plant\\_sexuality](http://en.wikipedia.org/wiki/Plant_sexuality)

## 5. Probability is in the Mind<sup>1</sup>

Monsterwith-  
girl\_2

**Followup to:** The Mind Projection Fallacy<sup>2</sup>

Yesterday I spoke of the Mind Projection Fallacy, giving the example of the alien monster who carries off a girl in a torn dress for intended ravishing—a mistake which I imputed to the artist’s tendency to think that a woman’s sexiness is a property of the woman herself, woman.sexiness, rather than something that exists in the mind of an observer, and probably wouldn’t exist in an alien mind.

The term “Mind Projection Fallacy” was coined by the late great Bayesian Master, E. T. Jaynes, as part of his long and hard-fought battle against the accursed frequentists. Jaynes was of the opinion that probabilities were in the mind, not in the environment—that probabilities express ignorance, states of partial information; and if I am ignorant of a phenomenon, that is a fact about my state of mind, not a fact about the phenomenon.

I cannot do justice to this ancient war in a few words—but the classic example of the argument runs thus:

You have a coin.

The coin is biased.

You don’t know which way it’s biased or how much it’s biased. Someone just told you, “The coin is biased” and that’s all they said. This is all the information you have, and the only information you have.

You draw the coin forth, flip it, and slap it down.

Now—before you remove your hand and look at the result—are you willing to say that you assign a 0.5 probability to the coin having come up heads?

The frequentist says, “No. Saying ‘probability 0.5’ means that the coin has an inherent propensity to come up heads as often as tails, so that if we flipped the coin infinitely many times, the ratio of heads to tails would approach 1:1. But we know that the coin is biased, so it can have any probability of coming up heads *except* 0.5.”

1. [http://lesswrong.com/lw/0j/probability\\_is\\_in\\_the\\_mind/](http://lesswrong.com/lw/0j/probability_is_in_the_mind/)

2. Page 685, ‘Mind Projection Fallacy’.

The Bayesian says, “Uncertainty exists in the map, not in the territory. In the real world, the coin has either come up heads, or come up tails. Any talk of ‘probability’ must refer to the *information* that I have about the coin—my state of partial ignorance and partial knowledge—not just the coin itself. Furthermore, I have all sorts of theorems showing that if I don’t treat my partial knowledge a certain way<sup>3</sup>, I’ll make stupid bets. If I’ve got to plan, I’ll plan for a 50/50 state of uncertainty, where I don’t weigh outcomes conditional on heads any more heavily in my mind than outcomes conditional on tails. You can call that number whatever you like, but it has to obey the probability laws on pain of stupidity. So I don’t have the slightest hesitation about calling my outcome-weighting a probability.”

I side with the Bayesians. You may have noticed that about me.

Even before a fair coin is tossed, the notion that it has an *inherent* 50% probability of coming up heads may be just plain wrong. Maybe you’re holding the coin in such a way that it’s just about guaranteed to come up heads, or tails, given the force at which you flip it, and the air currents around you. But, if you don’t know which way the coin is biased on this one occasion, so what?

I believe there was a lawsuit where someone alleged that the draft lottery was unfair, because the slips with names on them were not being mixed thoroughly enough; and the judge replied, “To whom is it unfair?”

To make the coinflip experiment repeatable, as frequentists are wont to demand, we could build an automated coinflipper, and verify that the results were 50% heads and 50% tails. But maybe a robot with extra-sensitive eyes and a good grasp of physics, watching the autoflipper prepare to flip, could predict the coin’s fall in advance—not with certainty, but with 90% accuracy. Then what would the *real* probability be?

There is no “real probability”. The robot has one state of partial information. You have a different state of partial information. The coin itself has no mind, and doesn’t assign a probability to anything; it just flips into the air, rotates a few times, bounces off some air molecules, and lands either heads or tails.

---

3. Page 644, ‘Something to Protect’.

So that is the Bayesian view of things, and I would now like to point out a couple of classic brainteasers that derive their brain-teasing ability from the tendency to think of probabilities as inherent properties of objects.

Let's take the old classic: You meet a mathematician on the street, and she happens to mention that she has given birth to two children on two separate occasions. You ask: "Is at least one of your children a boy?" The mathematician says, "Yes, he is."

What is the probability that she has two boys? If you assume that the prior probability of a child being a boy is  $1/2$ , then the probability that she has two boys, on the information given, is  $1/3$ . The prior probabilities were:  $1/4$  two boys,  $1/2$  one boy one girl,  $1/4$  two girls. The mathematician's "Yes" response has probability -1 in the first two cases, and probability -0 in the third. Renormalizing leaves us with a  $1/3$  probability of two boys, and a  $2/3$  probability of one boy one girl.

But suppose that instead you had asked, "Is your eldest child a boy?" and the mathematician had answered "Yes." Then the probability of the mathematician having two boys would be  $1/2$ . Since the eldest child is a boy, and the younger child can be anything it pleases.

Likewise if you'd asked "Is your youngest child a boy?" The probability of their being both boys would, again, be  $1/2$ .

Now, if at least one child is a boy, it must be either the oldest child who is a boy, or the youngest child who is a boy. So how can the answer in the first case be different from the answer in the latter two?

Or here's a very similar problem: Let's say I have four cards, the ace of hearts, the ace of spades, the two of hearts, and the two of spades. I draw two cards at random. You ask me, "Are you holding at least one ace?" and I reply "Yes." What is the probability that I am holding a pair of aces? It is  $1/5$ . There are six possible combinations of two cards, with equal prior probability, and you have just eliminated the possibility that I am holding a pair of twos. Of the five remaining combinations, only one combination is a pair of aces. So  $1/5$ .

Now suppose that instead you asked me, "Are you holding the ace of spades?" If I reply "Yes", the probability that the other card

is the ace of hearts is  $1/3$ . (You know I'm holding the ace of spades, and there are three possibilities for the other card, only one of which is the ace of hearts.) Likewise, if you ask me "Are you holding the ace of hearts?" and I reply "Yes", the probability I'm holding a pair of aces is  $1/3$ .

But then how can it be that if you ask me, "Are you holding at least one ace?" and I say "Yes", the probability I have a pair is  $1/5$ ? Either I must be holding the ace of spades or the ace of hearts, as you know; and either way, the probability that I'm holding a pair of aces is  $1/3$ .

How can this be? Have I miscalculated one or more of these probabilities?

If you want to figure it out for yourself, do so now, because I'm about to reveal...

That all stated calculations are correct.

As for the paradox, there isn't one. The *appearance* of paradox comes from thinking that the probabilities must be properties of the cards themselves. The ace I'm holding has to be either hearts or spades; but that doesn't mean that your *knowledge about* my cards must be the same as if you *knew* I was holding hearts, or *knew* I was holding spades.

It may help to think of Bayes's Theorem:

$$P(H|E) = P(E|H)P(H) / P(E)$$

That last term, where you divide by  $P(E)$ , is the part where you throw out all the possibilities that have been eliminated, and renormalize your probabilities over what remains.

Now let's say that you ask me, "Are you holding at least one ace?" *Before* I answer, your probability that I say "Yes" should be  $5/6$ .

But if you ask me "Are you holding the ace of spades?", your prior probability that I say "Yes" is just  $1/2$ .

So right away you can see that you're *learning* something very different in the two cases. You're going to be eliminating some different possibilities, and renormalizing using a different  $P(E)$ . If you learn two different items of evidence, you shouldn't be surprised at ending up in two different states of partial information.

Similarly, if I ask the mathematician, “Is at least one of your two children a boy?” I expect to hear “Yes” with probability  $3/4$ , but if I ask “Is your eldest child a boy?” I expect to hear “Yes” with probability  $1/2$ . So it shouldn’t be surprising that I end up in a different state of partial knowledge, depending on which of the two questions I ask.

The only reason for seeing a “paradox” is thinking as though the probability of holding a pair of aces is *a property of cards* that have at least one ace, or a property of cards that happen to contain the ace of spades. In which case, it would be paradoxical for card-sets containing at least one ace to have an inherent pair-probability of  $1/5$ , while card-sets containing the ace of spades had an inherent pair-probability of  $1/3$ , and card-sets containing the ace of hearts had an inherent pair-probability of  $1/3$ .

Similarly, if you think a  $1/3$  probability of being both boys is an *inherent property* of child-sets that include at least one boy, then that is not consistent with child-sets of which the eldest is male having an *inherent* probability of  $1/2$  of being both boys, and child-sets of which the youngest is male having an inherent  $1/2$  probability of being both boys. It would be like saying, “All green apples weigh a pound, and all red apples weigh a pound, and all apples that are green or red weigh half a pound.”

That’s what happens when you start thinking as if probabilities are *in* things, rather than probabilities being states of partial information *about* things.

Probabilities express uncertainty, and it is only agents who can be uncertain. A blank map does not correspond to a blank territory. Ignorance is in the mind.

## 6. The Quotation is not the Referent<sup>1</sup>

**Followup to:** The Mind Projection Fallacy<sup>2</sup>, Probability is in the Mind<sup>3</sup> <sup>4</sup>

In classical logic, the operational definition of identity is that whenever ‘A=B’ is a theorem, you can substitute ‘A’ for ‘B’ in any theorem where B appears. For example, if  $(2 + 2) = 4$  is a theorem, and  $((2 + 2) + 3) = 7$  is a theorem, then  $(4 + 3) = 7$  is a theorem.

This leads to a problem which is usually phrased in the following terms: The morning star and the evening star happen to be the same object, the planet Venus. Suppose John knows that the morning star and evening star are the same object. Mary, however, believes that the morning star is the god Lucifer, but the evening star is the god Venus. John believes Mary believes that the morning star is Lucifer. Must John therefore (by substitution) believe that Mary believes that the evening star is Lucifer?

Or here's an even simpler version of the problem.  $2 + 2 = 4$  is true; it is a theorem that  $((2 + 2) = 4) = \text{TRUE}$ . Fermat's Last Theorem is also true. So: I believe  $2 + 2 = 4 \Rightarrow$  I believe TRUE  $\Rightarrow$  I believe Fermat's Last Theorem.

Yes, I know this seems *obviously* wrong. But imagine someone writing a logical reasoning program using the principle “equal terms can always be substituted”, and this happening to them. Now imagine them writing a paper about how to prevent it from happening. Now imagine someone else disagreeing with their solution. The argument is still going on.

Personally, I would say that John is committing a type error, like trying to subtract 5 grams from 20 meters. “The morning star” is not the same *type* as the morning star, let alone the same thing. Beliefs are not planets.

morning star = evening star  
“morning star” ≠ “evening star”

---

1. [http://lesswrong.com/lw/ok/the\\_quotation\\_is\\_not\\_the\\_referent/](http://lesswrong.com/lw/ok/the_quotation_is_not_the_referent/)

2. Page 685, ‘Mind Projection Fallacy’.

3. Page 687, ‘Probability is in the Mind’.

4. Page 253, ‘Fallacies of Compression’.

The problem, in my view, stems from the failure to enforce the type distinction between beliefs and things. The original error was writing an AI that stores its beliefs about Mary's beliefs about "the morning star" using the same representation as in its beliefs about the morning star.

If Mary believes the "morning star" is Lucifer, that doesn't mean Mary believes the "evening star" is Lucifer, because "morning star" ≠ "evening star". The whole paradox stems from the failure to use quote marks in appropriate places.

You may recall that this is not the first time I've talked about enforcing type discipline—the last time was when I spoke about the error of confusing expected utilities with utilities<sup>5</sup>. It is immensely helpful, when one is first learning physics, to learn to keep track of one's units—it may seem like a bother to keep writing down 'cm' and 'kg' and so on, until you notice that (a) your answer seems to be the wrong order of magnitude and (b) it is expressed in seconds per square gram.

Similarly, beliefs are different things than planets. If we're talking about human beliefs, at least, then: Beliefs live in brains, planets live in space. Beliefs weigh a few micrograms, planets weigh a lot more. Planets are larger than beliefs... but you get the idea.

Merely putting quote marks around "morning star" seems insufficient to prevent people from confusing it with the morning star, due to the visual similarity of the text. So perhaps a better way to enforce type discipline would be with a visibly different encoding:

morning star = evening star

13.i5.i8.i4.9.i4.7.0.i9.20.i18 ≠ 5.22.5.i4.9.i4.7.0.i9.20.i18

Studying mathematical logic may also help you learn to distinguish the quote and the referent. In mathematical logic,  $\vdash P$  ( $P$  is a theorem) and  $\vdash [ ]' P'$  (it is provable that there exists an encoded proof of the encoded sentence  $P$  in some encoded proof system) are very distinct propositions. If you drop a level of quotation in mathematical logic, it's like dropping a metric unit in physics—you can derive visibly ridiculous results, like "The speed of light is 299,792,458 meters long."

---

<sup>5</sup>. [http://lesswrong.com/lw/l4/terminal\\_values\\_and\\_instrumental\\_values/](http://lesswrong.com/lw/l4/terminal_values_and_instrumental_values/)

Alfred Tarski once tried to define the meaning of ‘true’ using an infinite family of sentences:

(“Snow is white” is true) if and only (snow is white)  
 (“Weasels are green” is true) if and only if (weasels are  
 green)  
 ...

When sentences like these start seeming meaningful, you’ll know that you’ve started to distinguish between encoded sentences and states of the outside world.

Similarly, the notion of truth<sup>6</sup> is quite different from the notion of *reality*. Saying “true” *compares* a belief to reality. Reality itself does not need to be compared to any beliefs in order to be real. Remember this the next time someone claims that nothing is true.

---

6. <http://yudkowsky.net/bayes/truth.html>

## 7. Qualitatively Confused<sup>1</sup>

**Followup to:** Probability is in the Mind<sup>2</sup>, The Quotation is not the Referent<sup>3</sup>

I suggest that a primary cause of confusion about the distinction between “belief”, “truth”, and “reality” is qualitative thinking<sup>4</sup> about beliefs.

Consider the archetypal postmodernist attempt to be clever:

“The Sun goes around the Earth” is true for Hunga Huntergatherer, but “The Earth goes around the Sun” is true for Amara Astronomer! Different societies have different truths!

No, different societies have different *beliefs*. Belief is of a different type than truth; it’s like comparing apples and probabilities.

Ah, but there’s no difference between the way you use the word ‘belief’ and the way you use the word ‘truth’! Whether you say, “I believe ‘snow is white’”, or you say, “‘Snow is white’ is true”, you’re expressing exactly the same opinion.

No, these sentences mean quite different things, which is how I can *conceive* of the possibility that my beliefs are false.

Oh, you claim to *conceive* it, but you never *believe* it. As Wittgenstein said, “If there were a verb meaning ‘to believe falsely’, it would not have any significant first person, present indicative.”

And that’s what I mean by putting my finger on qualitative reasoning as the source of the problem. The dichotomy between belief and disbelief, being binary, is confusingly similar to the dichotomy between truth and untruth.

---

1. [http://lesswrong.com/lw/om/qualitatively\\_confused/](http://lesswrong.com/lw/om/qualitatively_confused/)

2. Page 687, ‘Probability is in the Mind’.

3. Page 692, ‘The Quotation is not the Referent’.

4. Page 599, ‘The Fallacy of Gray’.

So let's use quantitative reasoning instead. Suppose that I assign a 70% probability to the proposition that snow is white. It follows that I think there's around a 70% chance that the sentence "snow is white" will turn out to be true. If the sentence "snow is white" is true, is my 70% probability assignment to the proposition, also "true"? Well, it's more true than it would have been if I'd assigned 60% probability, but not so true as if I'd assigned 80% probability.

When talking about the correspondence between a probability assignment and reality, a better word than "truth" would be "accuracy". "Accuracy" sounds more quantitative, like an archer shooting an arrow: how close did your probability assignment strike to the center of the target?

To make a long story<sup>5</sup> short, it turns out that there's a very natural way of scoring the accuracy of a probability assignment, as compared to reality: just take the logarithm of the probability assigned to the real state of affairs.

So if snow is white, my belief "70%: 'snow is white'" will score<sup>6</sup> -0.51 bits:  $\text{Log}_2(0.7) = -0.51$ .

But what if snow is not white, as I have conceded a 30% probability is the case? If "snow is white" is false, my belief "30% probability: 'snow is not white'" will score -1.73 bits. Note that  $-1.73 < -0.51$ , so I have done worse.

About how accurate do I think my own beliefs are? Well, my expectation over the score is  $70\% * -0.51 + 30\% * -1.73 = -0.88$  bits. If snow is white, then my beliefs will be more accurate than I expected; and if snow is not white, my beliefs will be less accurate than I expected; but in neither case will my belief be *exactly* as accurate as I expected on average.

All this should not be confused with the statement "I assign 70% credence that 'snow is white'." I may well believe *that* proposition with probability -1—be quite certain that this is in fact my belief. If so I'll expect my meta-belief " $\neg I$ : 'I assign 70% credence that 'snow is white''' to score -0 bits of accuracy, which is as good as it gets.

---

5. <http://yudkowsky.net/bayes/technical.html>

6. <http://yudkowsky.net/bayes/technical.html>

Just because I am uncertain about snow, does not mean I am uncertain about my *quoted probabilistic beliefs*. Snow is out there, my beliefs are inside me. I may be a great deal less uncertain about how uncertain I am about snow, than I am uncertain about snow. (Though beliefs about beliefs are not always accurate<sup>7</sup>.)

Contrast this probabilistic situation to the qualitative reasoning where I just believe that snow is white, and believe that I believe that snow is white, and believe ““snow is white” is true”, and believe “my belief ““snow is white” is true” is correct”, etc. Since all the quantities involved are 1, it’s easy to mix them up.

Yet the nice distinctions of quantitative reasoning will be short-circuited if you start thinking ““snow is white” with 70% probability’ is *true*“, which is a type error. It is a true fact about you, that you *believe* “70% probability: ‘snow is white’”; but that does not mean the probability assignment *itself* can possibly be “true”. The belief scores either -0.51 bits or -1.73 bits of accuracy, depending on the actual state of reality.

The cognoscenti will recognize ““snow is white” with 70% probability’ is *true*” as the mistake of thinking that probabilities are inherent properties of things<sup>8</sup>.

From the inside<sup>9</sup>, our beliefs about the world look like the world, and our beliefs about our beliefs look like beliefs. When you see the world, you are experiencing a belief from the inside. When you notice yourself believing something, you are experiencing a belief about belief from the inside. So if your internal representations of belief, and belief about belief, are dissimilar<sup>10</sup>, then you are less likely to mix them up and commit the Mind Projection Fallacy<sup>11</sup>—I hope.

When you think in probabilities, your beliefs, and your beliefs about your beliefs, will hopefully not be represented similarly enough that you mix up belief and accuracy, or mix up accuracy and reality. When you think in probabilities *about the world*, your beliefs will be represented with probabilities  $\in (0, 1)$ . Unlike the truth-values of propositions, which are in {true, false}. As for the accuracy

7. Page 108, ‘Belief in Belief’.

8. Page 687, ‘Probability is in the Mind’.

9. Page 223, ‘How An Algorithm Feels From Inside’.

10. Page 692, ‘The Quotation is not the Referent’.

11. Page 685, ‘Mind Projection Fallacy’.

of your probabilistic belief, you can represent that in the range  $(-\infty, 0)$ . Your probabilities *about your beliefs* will typically be extreme. And things themselves—why, they’re just red, or blue, or weighing 20 pounds, or whatever.

Thus we will be less likely, perhaps, to mix up the map with the territory.

This type distinction may also help us remember that *uncertainty* is a state of mind. A coin is not *inherently* 50% uncertain of which way it will land. The coin is not a belief processor, and does not have partial information about itself. In qualitative reasoning you can create a belief that corresponds very straightforwardly to the coin, like “The coin will land heads”. This belief will be true or false *depending on* the coin, and there will be a transparent implication from the truth or falsity of the belief, to the facing side of the coin.

But even under qualitative reasoning, to say that the coin *itself* is “true” or “false” would be a severe type error. The coin is not a belief, it is a coin. The territory is not the map.

If a coin cannot be true or false, how much less can it assign a 50% probability to itself?

## 8. Reductionism<sup>1</sup>

**Followup to:** How An Algorithm Feels From Inside<sup>2</sup>, Mind Projection Fallacy<sup>3</sup>

Almost one year ago, in April 2007, Matthew C submitted the following suggestion for an Overcoming Bias topic:

“How and why the current reigning philosophical hegemon (reductionistic materialism) is obviously correct [...], while the reigning philosophical viewpoints of all past societies and civilizations are obviously suspect—”

I remember this, because I looked at the request and deemed it legitimate, but I knew I couldn’t do that topic until I’d started on the Mind Projection Fallacy<sup>4</sup> sequence, which wouldn’t be for a while...

But now it’s time to begin addressing this question. And while I haven’t yet come to the “materialism” issue, we can now start on “reductionism”.

First, let it be said that I do indeed hold that “reductionism”, according to the meaning<sup>5</sup> I will give for that word, is obviously correct; and to perdition with any past civilizations that disagreed<sup>6</sup>.

This seems like a strong statement, at least the first part of it. General Relativity seems well-supported, yet who knows but that some future physicist may overturn it?

On the other hand, we are never going *back* to Newtonian mechanics. The ratchet of science turns, but it does not turn in reverse. There are cases in scientific history where a theory suffered a wound or two, and then bounced back; but when a theory takes as many arrows through the chest as Newtonian mechanics, it *stays dead*.

---

1. <http://lesswrong.com/lw/on/reductionism/>

2. Page 223, ‘How An Algorithm Feels From Inside’.

3. Page 685, ‘Mind Projection Fallacy’.

4. Page 685, ‘Mind Projection Fallacy’.

5. Page 238, ‘The Argument from Common Usage’.

6. Page 409, ‘Guardians of the Truth’.

“To hell with what past civilizations thought” seems safe enough, when past civilizations believed in something that has been falsified to the trash heap of history.

And reductionism is not so much a positive hypothesis, as the *absence* of belief—in particular, disbelief in a form of the Mind Projection Fallacy.

I once met a fellow who claimed that he had experience as a Navy gunner, and he said, “When you fire artillery shells, you’ve got to compute the trajectories using Newtonian mechanics. If you compute the trajectories using relativity, you’ll get the wrong answer.”

And I, and another person who was present, said flatly, “No.” I added, “You might not be able to compute the trajectories fast enough to get the answers in time—maybe that’s what you mean? But the relativistic answer will always be more accurate than the Newtonian one.”

“No,” he said, “I mean that relativity will give you the *wrong answer*, because things moving at the speed of artillery shells are governed by Newtonian mechanics, not relativity.”

“If that were really true,” I replied, “you could publish it in a physics journal and collect your Nobel Prize.”

Standard physics uses the same *fundamental* theory to describe the flight of a Boeing 747 airplane, and collisions in the Relativistic Heavy Ion Collider. Nuclei and airplanes alike, according to our understanding, are obeying special relativity, quantum mechanics, and chromodynamics.

But we use entirely different *models* to understand the aerodynamics of a 747 and a collision between gold nuclei in the RHIC. A computer modeling the aerodynamics of a 747 may not contain a single token, a single bit of RAM, that represents a quark.

So is the 747 made of something other than quarks? No, you’re just *modeling* it with *representational elements* that do not have a one-to-one correspondence with the quarks of the 747. The map is not the territory.

Why *not* model the 747 with a chromodynamic representation? Because then it would take a gazillion years to get any answers out of the model. Also we could not store the model on all the memory on all the computers in the world, as of 2008.

As the saying goes, “The map is not the territory, but you can’t fold up the territory and put it in your glove compartment.” Sometimes you need a smaller map to fit in a more cramped glove compartment—but this does not change the territory. The scale of a map is not a fact about the territory, it’s a fact about the map.

If it *were* possible to build and run a chromodynamic model of the 747, it would yield accurate predictions. Better predictions than the aerodynamic model, in fact.

To build a fully accurate model of the 747, it is not necessary, in principle, for the model to contain explicit descriptions of things like airflow and lift. There does not have to be a single token, a single bit of RAM, that corresponds to the position of the wings. It is possible, in principle, to build an accurate model of the 747 that makes no mention of anything *except* elementary particle fields and fundamental forces.

“What?” cries the antireductionist. “Are you telling me the 747 *doesn’t really have wings?* I can see the wings right there!”

The notion here is a subtle one. It’s not *just* the notion that an object can have different descriptions at different levels.

It’s the notion that “having different descriptions at different levels” is *itself* something you say that belongs in the realm of Talking About Maps, not the realm of Talking About Territory.

It’s not that the *airplane itself*, the *laws of physics themselves*, use different descriptions at different levels—as yonder artillery gunner thought. Rather *we*, for our convenience, use different simplified models at different levels.

If you looked at the ultimate chromodynamic model, the one that contained only elementary particle fields and fundamental forces, that model would contain all the facts about airflow and lift and wing positions—but these facts would be implicit, rather than explicit.

You, looking *at* the model, and thinking *about* the model, would be able to figure out where the wings were. Having figured it out, there would be an explicit representation in your mind of the wing position—an explicit computational object, there in your neural RAM. *In your mind.*

You might, indeed, deduce all sorts of explicit descriptions of the airplane, at various levels, and even explicit rules for how your

models at different levels interacted with each other to produce combined predictions—

And the way that algorithm feels from inside<sup>7</sup>, is that the airplane would *seem* to be made up of many levels at once, interacting with each other.

The way a belief *feels from inside*, is that you seem to be looking straight at reality. When it actually *seems* that you're looking at a belief, as such, you are really experiencing a belief about belief<sup>8</sup>.

So when your mind simultaneously believes explicit descriptions of many different levels, and believes explicit rules for transiting between levels, as part of an efficient combined model, it *feels like* you are seeing a system that is *made of* different level descriptions and their rules for interaction.

But this is just the brain trying to be efficiently compress an object that it cannot remotely begin to model on a fundamental level. The airplane is too large. Even a hydrogen atom would be too large. Quark-to-quark interactions are insanely intractable. You can't handle the *truth*.

But the way physics *really* works, as far as we can tell, is that there is *only* the most basic level—the elementary particle fields and fundamental forces. You can't handle the raw truth, but reality can handle it without the slightest simplification. (I wish I knew where Reality got its computing power.)

The laws of physics do not contain distinct additional causal entities that correspond to lift or airplane wings, the way that *the mind of an engineer* contains distinct additional *cognitive* entities that correspond to lift or airplane wings.

This, as I see it, is the thesis of reductionism. Reductionism is not a positive belief, but rather, a disbelief that the higher levels of simplified multilevel models are out there in the territory. Understanding this on a gut level dissolves the question<sup>9</sup> of “How can you say the airplane doesn't really have wings, when I can see the wings right there?” The critical words are *really* and *see*.

7. Page 223, ‘How An Algorithm Feels From Inside’.

8. Page 695, ‘Qualitatively Confused’.

9. Page 673, ‘Dissolving the Question’.

## 9. Explaining vs. Explaining Away<sup>1</sup>

**Followup to:** Reductionism<sup>2</sup>, Righting a Wrong Question<sup>3</sup>

John Keats's *Lamia*<sup>4</sup> (1819) surely deserves some kind of award for Most Famously Annoying Poetry:

...Do not all charms fly  
At the mere touch of cold philosophy?  
There was an awful rainbow once in heaven:  
We know her woof, her texture; she is given  
In the dull catalogue of common things.  
Philosophy will clip an Angel's wings,  
Conquer all mysteries by rule and line,  
Empty the haunted air, and gnomed mine—  
Unweave a rainbow...

My usual reply ends with the phrase: “If we cannot learn to take joy in the merely real, our lives will be empty indeed.” I shall expand on that tomorrow.

Today I have a different point in mind. Let’s just take the lines:

Empty the haunted air, and gnomed mine—  
Unweave a rainbow...

Apparently “the mere touch of cold philosophy”, i.e., the truth, has destroyed:

- Haunts in the air
- Gnomes in the mine
- Rainbows

Which calls to mind a rather different bit of verse<sup>5</sup>:

One of these things  
Is not like the others  
One of these things  
Doesn’t belong

1. [http://lesswrong.com/lw/oo/explaining\\_vs\\_explaining\\_away/](http://lesswrong.com/lw/oo/explaining_vs_explaining_away/)

2. Page 699, ‘Reductionism’.

3. Page 682, ‘Righting a Wrong Question’.

4. <http://en.wikisource.org/wiki/Lamia>

5. <http://www.youtube.com/watch?v=Ect-kgxBb4M>

The air has been emptied of its haunts, and the mine de-gnomed—but the rainbow is still there!

In “Righting a Wrong Question<sup>6</sup>”, I wrote:

Tracing back the chain of causality, step by step, I discover that my belief that I’m wearing socks is fully explained by the fact that I’m wearing socks... On the other hand, if I see a mirage of a lake in the desert, the correct causal explanation of my vision does not involve the fact of any actual lake in the desert. In this case, my belief in the lake is not just *explained*, but *explained away*.

The rainbow was *explained*. The haunts in the air, and gnomes in the mine, were *explained away*.

I think this is the key distinction that anti-reductionists don’t get about reductionism.

You can see this failure to get the distinction in the classic objection to reductionism:

If reductionism is correct, then even your belief in reductionism is just the mere result of the motion of molecules—why should I listen to anything you say?

The key word, in the above, is *mere*; a word which implies that accepting reductionism would explain *away* all the reasoning processes leading up to my acceptance of reductionism, the way that an optical illusion is explained *away*.

But you can explain how a cognitive process works without it being “mere”! My belief that I’m wearing socks is a mere result of my visual cortex reconstructing nerve impulses sent from my retina which received photons reflected off my socks... which is to say, according to scientific reductionism, my belief that I’m wearing socks is a mere result of the fact that I’m wearing socks.

What could be going on in the anti-reductionists’ minds<sup>7</sup>, such that they would put rainbows and belief-in-reductionism, in the same category as haunts and gnomes?

---

6. Page 682, ‘Righting a Wrong Question’.

7. Page 673, ‘Dissolving the Question’.

Several things are going on simultaneously. But for now let's focus on the basic idea introduced yesterday: The Mind Projection Fallacy<sup>8</sup> between a multi-level map and a mono-level territory.

(I.e: There's no way you can model a 747 quark-by-quark, so you've *got* to use a multi-level map with explicit cognitive representations of wings, airflow, and so on. This doesn't mean there's a multi-level territory. The true laws of physics, to the best of our knowledge, are only over elementary particle fields.)

I think that when physicists say "There are no *fundamental* rainbows," the anti-reductionists hear, "There are no rainbows."

If you don't distinguish between the multi-level map and the mono-level territory, then when someone tries to explain to you that the rainbow is not a fundamental thing in physics, acceptance of this will *feel like* erasing rainbows from your multi-level map, which *feels like* erasing rainbows from the world.

When Science says "tigers are not *elementary* particles, they are made of quarks" the anti-reductionist hears this as the same sort of dismissal as "we looked in your garage for a dragon, but there was just empty air".

What scientists did to rainbows, and what scientists did to gnomes, seemingly felt the same to Keats...

In support of this sub-thesis, I deliberately used several phrasings, in my discussion of Keats's poem, that were Mind Projection Fallacious. If you didn't notice, this would seem to argue that such fallacies are customary enough to pass unremarked.

For example:

"The air has been emptied of its haunts, and the mine de-gnomed—but the rainbow is still there!"

Actually, Science emptied the *model of* air of *belief in* haunts, and emptied the *map of* the mine of *representations of* gnomes. Science did not actually—as Keats's poem itself would have it—take real Angel's wings, and destroy them with a cold touch of truth. In reality there *never were* any haunts in the air, or gnomes in the mine.

Another example:

---

8. Page 685, 'Mind Projection Fallacy'.

“What scientists did to rainbows, and what scientists did to gnomes, seemingly felt the same to Keats.”

Scientists didn’t *do* anything *to* gnomes, only to “gnomes”. The quotation is not the referent.

But if you commit the Mind Projection Fallacy—and by default, our beliefs just feel like the way the world *is*—then at time T=0, the mines (apparently) contain gnomes; at time T=1 a scientist dances across the scene, and at time T=2 the mines (apparently) are empty. Clearly, there used to be gnomes there, but the scientist killed them.

Bad scientist! No poems for you, gnomekiller!

Well, that’s how it *feels*, if you get emotionally attached to the gnomes, and then a scientist says there aren’t any gnomes. It takes a strong mind, a deep honesty, and a deliberate effort to say, at this point, “That which can be destroyed by the truth should be,” and “The scientist hasn’t taken the gnomes away, only taken my delusion away,” and “I never held just title to my belief in gnomes in the first place; I have not been deprived of anything I *rightfully* owned,” and “If there are gnomes, I desire to believe there are gnomes; if there are no gnomes, I desire to believe there are no gnomes; let me not become attached to beliefs I may not want,” and all the other things that rationalists are supposed to say on such occasions.

But with the rainbow it is not even necessary to go that far. The rainbow is *still there!*

## 10. Fake Reductionism<sup>1</sup>

**Followup to:** Explaining vs. Explaining Away<sup>2</sup>, Fake Explanation<sup>3</sup>

There was an awful rainbow once in heaven:  
We know her woof, her texture; she is given  
In the dull catalogue of common things.

—John Keats, *Lamia*

I am guessing—though it is only a guess—that Keats himself did *not* know the woof and texture of the rainbow. Not the way that Newton understood rainbows. Perhaps not even at all. Maybe Keats just read, somewhere, that Newton had explained the rainbow as “light reflected from raindrops”—

—which was actually known in the 13th century. Newton only added a refinement by showing that the light was decomposed into colored parts, rather than transformed in color. But that put rainbows back in the news headlines. And so Keats, with Charles Lamb and William Wordsworth and Benjamin Haydon, drank “Confusion to the memory of Newton” because “he destroyed the poetry of the rainbow by reducing it to a prism.” That’s one reason to suspect Keats didn’t understand the subject too deeply.

I am guessing, though it is only a guess, that Keats could *not* have sketched out on paper why rainbows only appear when the Sun is behind your head, or why the rainbow is an arc of a circle.

If so, Keats had a Fake Explanation<sup>4</sup>. In this case, a *fake reduction*. He’d been *told that* the rainbow had been reduced, but it had not actually *been reduced* in his model of the world.

This is another of those distinctions that anti-reductionists fail to get—the difference between professing<sup>5</sup> the flat fact that something is reducible, and *seeing* it.

In this, the anti-reductionists are not too greatly to be blamed, for it is part of a general problem.

---

1. [http://lesswrong.com/lw/op/fake\\_reductionism/](http://lesswrong.com/lw/op/fake_reductionism/)

2. Page 703, ‘Explaining vs. Explaining Away’.

3. Page 138, ‘Fake Explanations’.

4. Page 138, ‘Fake Explanations’.

5. Page 108, ‘Belief in Belief’.

I've written before on seeming knowledge that is not knowledge<sup>6</sup>, and beliefs that are not *about* their supposed objects but only recordings to recite back in the classroom<sup>7</sup>, and words that operate as stop signs for curiosity<sup>8</sup> rather than answers, and technobabble which only conveys membership in the literary genre of "science"<sup>9</sup>...

There is a very great distinction between being able to *see* where the rainbow comes from, and playing around with prisms to confirm it, and maybe making a rainbow yourself by spraying water droplets—

—versus some dour-faced philosopher just *telling* you, "No, there's nothing special about the rainbow. Didn't you hear? Scientists have explained it away. Just something to do with raindrops or whatever. Nothing to be excited about."

I think this distinction probably accounts for a hell of a lot of the deadly existential emptiness that supposedly accompanies scientific reductionism.

You have to interpret the anti-reductionists' experience of "reductionism", not in terms of their *actually seeing* how rainbows work, not in terms of their having the critical "Aha!", but in terms of their being told that the password<sup>10</sup> is "Science". The effect is just to move rainbows to a different *literary genre*—a literary genre<sup>11</sup> they have been taught<sup>12</sup> to regard as boring<sup>13</sup>.

For them, the effect of hearing "Science has explained rainbows!" is to hang up a sign over rainbows saying, "This phenomenon has been labeled BORING by order of the Council of Sophisticated Literary Critics. Move along."

And that's all the sign says: only that, and nothing more.

So the literary critics have their gnomes yanked out by force; not dissolved in insight, but removed by flat order of authority. They are given no beauty to replace the hauntings air, no genuine understanding that could be interesting in its own right. Just a label

6. Page 138, 'Fake Explanations'.

7. Page 140, 'Guessing the Teacher's Password'.

8. Page 151, 'Semantic Stopsigns'.

9. Page 144, 'Science as Attire'.

10. Page 140, 'Guessing the Teacher's Password'.

11. Page 144, 'Science as Attire'.

12. Page 451, 'Cached Thoughts'.

13. Page 177, "Science" as Curiosity-Stopper'.

saying, “Ha! You thought rainbows were pretty? You poor, unsophisticated fool. This is part of the literary genre of science, of dry and solemn incomprehensible words.”

That’s how anti-reductionists experience “reductionism”.

Well, can’t blame Keats, poor lad probably wasn’t raised right.

But he dared to drink “Confusion to the memory of Newton”?

I propose “To the memory of Keats’s confusion” as a toast for rationalists. Cheers.

## II. Savanna Poets<sup>1</sup>

**Followup to:** Explaining vs. Explaining Away<sup>2</sup>

“Poets say science takes away from the beauty of the stars—mere globs of gas atoms. Nothing is “mere”. I too can see the stars on a desert night, and feel them. But do I see less or more?

“The vastness of the heavens stretches my imagination—stuck on this carousel my little eye can catch one-million-year-old light. A vast pattern—of which I am a part—perhaps my stuff was belched from some forgotten star, as one is belching there. Or see them with the greater eye of Palomar, rushing all apart from some common starting point when they were perhaps all together. What is the pattern, or the meaning, or the why? It does not do harm to the mystery to know a little about it.

“For far more marvelous is the truth than any artists of the past imagined! Why do the poets of the present not speak of it?

“What men are poets who can speak of Jupiter if he were like a man, but if he is an immense spinning sphere of methane and ammonia must be silent?”

—Richard Feynman, *The Feynman Lectures on Physics*, Vol I, p. 3-6 (line breaks added)

That’s a real question, there on the last line—what kind of poet can write about Jupiter the god, but not Jupiter the immense sphere? Whether or not Feynman meant the question rhetorically, it has a real answer:

If Jupiter is like us, he can fall in love, and lose love, and regain love.

If Jupiter is like us, he can strive, and rise, and be cast down.

If Jupiter is like us, he can laugh or weep or dance.

If Jupiter is an immense spinning sphere of methane and ammonia, it is more difficult for the poet to make us feel.

---

1. [http://lesswrong.com/lw/oq/savanna\\_poets/](http://lesswrong.com/lw/oq/savanna_poets/)

2. Page 703, ‘Explaining vs. Explaining Away’.

There are poets and storytellers who say that the Great Stories are timeless, and they never change, they only ever retold. They say, with pride, that Shakespeare and Sophocles are bound by ties of craft stronger than mere centuries; that the two playwrights could have swapped times without a jolt.

Donald Brown once compiled a list of over two hundred “human universals<sup>3</sup>”, found in all (or a vast supermajority of) studied human cultures, from San Francisco to the !Kung of the Kalahari Desert. Marriage is on the list, and incest avoidance, and motherly love, and sibling rivalry, and music and envy and dance and storytelling and aesthetics, and ritual magic to heal the sick, and poetry in spoken lines separated by pauses—

No one who knows anything about evolutionary psychology<sup>4</sup> could be expected to deny it: The strongest emotions we have are deeply engraved, blood and bone, brain and DNA.

It might take a bit of tweaking, but you probably *could* tell “Hamlet” sitting around a campfire on the ancestral savanna.

So one can see why John “Unweave a rainbow” Keats might feel something had been lost, on being told that the rainbow was sunlight scattered from raindrops. Raindrops don’t dance.

In the Old Testament, it is written that God once destroyed the world with a flood that covered all the land, drowning all the horribly guilty men and women of the world along with their horribly guilty babies, but Noah built a gigantic wooden ark, etc., and after most of the human species was wiped out, God put rainbows in the sky as a sign that he wouldn’t do it again. At least not with water.

You can see how Keats would be *shocked* that this beautiful story was contradicted by modern science. Especially if (as I described yesterday<sup>5</sup>) Keats had no real understanding of rainbows, no “Aha!” insight that could be fascinating in its own right, to replace the drama subtracted—

Ah, but maybe Keats would be right to be disappointed *even if* he knew the math. The Biblical story of the rainbow is a tale of bloodthirsty murder and smiling insanity. How could anything

---

3. <http://condor.depaul.edu/~mfiddler/byphen/bumunivers.htm>

4. [http://lesswrong.com/lw/lx/evolutionary\\_psychology/](http://lesswrong.com/lw/lx/evolutionary_psychology/)

5. Page 707, ‘Fake Reductionism’.

about raindrops and refraction properly replace that? Raindrops don't scream when they die.

So science takes the romance away (says the Romantic poet), and what you are given back, never matches the drama of the original—

(that is, the original delusion<sup>6</sup>)

—even if you do know the equations, because the equations are not about strong emotions.

That is the strongest rejoinder I can think of, that any Romantic poet could have said to Feynman—though I can't remember ever hearing it said.

You can guess that I don't agree with the Romantic poets. So my own stance is this:

It is not *necessary* for Jupiter to be like a human, because *humans* are like humans. If Jupiter is an immense spinning sphere of methane and ammonia, that doesn't mean that love and hate are emptied from the universe. There *are* still loving and hating minds in the universe. *Us*.

With more than six billion of us at the last count, does Jupiter really need to be on the list of potential protagonists?

It is not *necessary* to tell the Great Stories about planets or rainbows. They play out all over our world, every day. Every day, someone kills for revenge; every day, someone kills a friend by mistake; every day, upward of a hundred thousand people fall in love. And even if this were not so, you could write fiction about humans—not about Jupiter.

Earth is old, and has played out the same stories many times beneath the Sun. I do wonder if it might not be time for some of the Great Stories to change. For me, at least, the story called “Goodbye”<sup>7</sup> has lost its charm.

The Great Stories are not timeless, because the human species is not timeless. Go far enough back in hominid evolution, and no one will understand *Hamlet*. Go far enough back in time, and you won't find any brains.

The Great Stories are not eternal, because the human species, *Homo sapiens sapiens*, is not eternal. I most sincerely doubt that we

6. Page 703, ‘Explaining vs. Explaining Away’.

7. <http://yudkowsky.net/yebuda.html>

have another thousand years to go in our current form. I do not say this in sadness: I think we can do better<sup>8</sup>.

I would not like to see all the Great Stories lost completely, in our future. I see very little difference between that outcome, and the Sun falling into a black hole.

But the Great Stories in their current forms have *already been told*, over and over. I do not think it ill if some of them should change their forms, or diversify their endings.

“And they lived happily ever after” seems worth trying at least once.

The Great Stories can and should diversify, as humankind grows up. Part of that ethic is the idea that when we find strangeness, we should respect it enough to tell its story truly. Even if it makes writing poetry a little more difficult.

If you are a good enough poet to write an ode to an immense spinning sphere of methane and ammonia, you are writing something *original*, about a newly discovered part of the real universe. It may not be as dramatic, or as gripping, as Hamlet. But the tale of Hamlet has already been told! If you write of Jupiter as though it were a human, then you are making our map of the universe just a little more impoverished of complexity; you are forcing Jupiter into the mold of all the stories that have already been told of Earth.

James Thomson’s “A Poem Sacred to the Memory of Sir Isaac Newton<sup>9</sup>”, which praises the rainbow for what it *really* is—you can argue whether or not Thomson’s poem is as gripping as John Keats’s Lamia<sup>10</sup> who was loved and lost. But tales of love and loss and cynicism had *already been told*, far away in ancient Greece, and no doubt many times before. Until we understood the rainbow as a thing *different* from tales of human-shaped magic, the true story of the rainbow could not be poeticized.

The border between science fiction and space opera was once drawn as follows: If you can take the plot of a story and put it back in the Old West, or the Middle Ages, without changing it, then it is not *real* science fiction. In real science fiction, the science is in-

---

8. [http://lesswrong.com/lw/b8/tsuyoku\\_naritai\\_i\\_want\\_to\\_become\\_stronger/](http://lesswrong.com/lw/b8/tsuyoku_naritai_i_want_to_become_stronger/)

9. <http://www.poembuilder.com/poem/a-poem-sacred-to-the-memory-of-sir-isaac-newton/>

10. <http://en.wikisource.org/wiki/Lamia>

trinsically part of the plot—you can't move the story from space to the savanna, not without losing something.

Richard Feynman asked: “What men are poets who can speak of Jupiter if he were like a man, but if he is an immense spinning sphere of methane and ammonia must be silent?”

They are *savanna poets*, who can *only* tell stories that would have made sense around a campfire ten thousand years ago. Savanna poets, who can tell *only* the Great Stories in their classic forms, and nothing more.

## 12. Joy in the Merely Real<sup>1</sup>

**Followup to:** Explaining vs. Explaining Away<sup>2</sup> <sup>3</sup>

...Do not all charms fly  
At the mere touch of cold philosophy?  
There was an awful rainbow once in heaven:  
We know her woof, her texture; she is given  
In the dull catalogue of common things.

—John Keats, *Lamia*

“Nothing is ‘mere’.”  
—Richard Feynman

You've got to admire that phrase, “dull catalogue of common things”. What is it, exactly, that goes in this catalogue? Besides rainbows, that is?

Why, things that are mundane, of course. Things that are normal; things that are unmagical; things that are known, or knowable; things that play by the rules (or that play by *any* rules, which makes them boring); things that are part of the ordinary universe; things that are, in a word, *real*.

Now that's what I call setting yourself up for a fall.

At that rate, sooner or later you're going to be disappointed in *everything*—either it will turn out not to exist, or even worse, it will turn out to be real.

If we cannot take joy in things that are merely real, our lives will *always* be empty.

For what sin are rainbows demoted to the dull catalogue of common things? For the sin of having a scientific explanation. “We know her woof, her texture”, says Keats—an interesting use of the word “we”, because I suspect that Keats didn't know<sup>4</sup> the explanation himself. I suspect that just being told that someone else knew<sup>5</sup> was too much for him to take. I suspect that just the notion of

---

1. [http://lesswrong.com/lw/or/joy\\_in\\_the\\_merely\\_real/](http://lesswrong.com/lw/or/joy_in_the_merely_real/)

2. Page 703, ‘Explaining vs. Explaining Away’.

3. Page 177, “Science” as Curiosity-Stopper.

4. Page 707, ‘Fake Reductionism’.

5. Page 177, “Science” as Curiosity-Stopper’.

rainbows being scientifically explicable *in principle* would have been too much to take. And if Keats didn't think like that, well, I know plenty of people who do.

I have already remarked that nothing is *inherently* mysterious<sup>6</sup>—nothing that actually exists, that is. If I am ignorant<sup>7</sup> about a phenomenon, that is a fact about my state of mind<sup>8</sup>, not a fact about the phenomenon; to worship<sup>9</sup> a phenomenon because it seems so wonderfully mysterious, is to worship your own ignorance; a blank map does not correspond to a blank territory, it is just somewhere we haven't visited yet, etc. etc...

Which is to say that *everything*—everything that *actually* exists—is liable to end up in “the dull catalogue of common things”, sooner or later.

Your choice is either:

- Decide that things are allowed to be unmagical, knowable, scientifically explicable, in a word, *real*, and yet still worth caring about;
- Or go about the rest of your life suffering from existential ennui that is *unresolvable*.

(Self-deception might be an option for others, but not for you<sup>10</sup>.)

This puts quite a different complexion on the bizarre habit indulged by those strange folk called *scientists*, wherein they suddenly become fascinated by pocket lint or bird droppings or rainbows, or some other ordinary thing which world-weary and sophisticated folk would never give a second glance.

You might say that scientists—at least *some* scientists—are those folk who are *in principle* capable of enjoying life in the real universe.

---

6. Page 154, ‘Mysterious Answers to Mysterious Questions’.

7. Page 687, ‘Probability is in the Mind’.

8. Page 685, ‘Mind Projection Fallacy’.

9. Page 175, ‘Explain/Worship/Ignore?’.

10. Page 565, ‘Doublethink (Choosing to be Biased)’.

## 13. Joy in Discovery<sup>1</sup>

**Followup to:** Joy in the Merely Real<sup>2</sup>

“Newton was the greatest genius who ever lived, and the most fortunate; for we cannot find more than once a system of the world to establish.”

—Lagrange

I have more fun discovering things for myself than reading about them in textbooks. This is right and proper, and only to be expected.

But discovering something that *no one else knows*—being the *first* to unravel the secret—

There is a story that one of the first men to realize that stars were burning by fusion—plausible attributions I've seen are to Fritz Houtermans<sup>3</sup> and Hans Bethe<sup>4</sup>—was walking out with his girlfriend of a night, and she made a comment on how beautiful the stars were, and he replied: “Yes, and right now, I'm the only man in the world who knows why they shine.”

It is attested by numerous sources that this experience, being the first person to solve a major mystery, is a *tremendous* high. It's probably the closest experience you can get to taking drugs, without taking drugs—though I wouldn't know.

*That* can't be healthy.

Not that I'm objecting to the euphoria. It's the exclusivity clause that bothers me. Why should a discovery be worth *less*, just because someone *else* already knows the answer?

The most charitable interpretation I can put on the psychology, is that you don't struggle with a single problem for months or years if it's something you can just look up in the library. And that the tremendous high comes from having hit the problem from every angle you can manage, and having bounced; and then having analyzed the problem again, using every idea you can think of, and all the data you can get your hands on—making progress a little at a time—so

---

1. [http://lesswrong.com/lw/os/joy\\_in\\_discovery/](http://lesswrong.com/lw/os/joy_in_discovery/)

2. Page 715, ‘Joy in the Merely Real’.

3. <http://ask.metafilter.com/54496/Feynman-Quote-Demystification>

4. <http://www.dykestowatchoutfor.com/more-on-consciousness#comment-72828>

that when, *finally*, you crack through the problem, all the dangling pieces and unresolved questions fall into place at once, like solving a dozen locked-room murder mysteries with a single clue.

And more, the understanding you get is *real* understanding—understanding that embraces all the clues you studied to solve the problem, when you didn’t yet know the answer. Understanding that comes from asking questions day after day and worrying at them; understanding that no one else can get (no matter how much you tell them the answer) unless they spend months studying the problem in its historical context, even after it’s been solved—and even then, they won’t get the high of solving it all at once.

That’s one possible reason why James Clerk Maxwell might have had more fun *discovering* Maxwell’s Equations, than you had fun reading about them.

A slightly less charitable reading is that the tremendous high comes from what is termed, in the *politesse* of social psychology, “commitment” and “consistency” and “cognitive dissonance”; the part where we value something more highly *just* because it took more work to get it. The studies showing that subjective fraternity pledges to a harsher initiation, causes them to be more convinced of the value of the fraternity—identical wine in higher-priced bottles being rated as tasting better—that sort of thing.

Of course, if you just have more fun solving a puzzle than being told its answer, because you enjoy doing the cognitive work for its own sake, there’s nothing wrong with that. The less charitable reading would be if charging \$100 to be told the answer to a puzzle, made you think the answer was more interesting, worthwhile, important, surprising, etc. than if you got the answer for free.

(I strongly suspect that a major part of science’s PR problem in the population at large is people who instinctively believe that if knowledge is given away for free, it cannot be important. If you had to undergo a fearsome initiation ritual to be told the truth about evolution, maybe people would be more satisfied with the answer.)

The really uncharitable reading is that the joy of first discovery is about status. Competition. Scarcity. Beating everyone else to the punch. It doesn’t matter whether you have a 3-room house or a 4-room house, what matters is having a bigger house than the Jone-

ses. A 2-room house would be fine, if you could only ensure that the Joneses had even less.

I don't object to competition as a matter of principle. I don't think that the game of Go is barbaric and should be suppressed, even though it's zero-sum. But if the euphoric joy of scientific discovery *has* to be about scarcity, that means it's only available to one person per civilization for any given truth.

If the joy of scientific discovery is one-shot per discovery, then, from a fun-theoretic perspective, Newton probably used up a substantial increment of the total Physics Fun available over the entire history of Earth-originating intelligent life. That selfish bastard explained the orbits of planets *and* the tides.

And really the situation is even worse than this, because in the Standard Model of physics (discovered by bastards who spoiled the puzzle for everyone else) the universe is spatially infinite, inflationarily branching, and branching via decoherence, which is at least three different ways that Reality is exponentially or infinitely large

So aliens, or alternate Newtons, or just Tegmark duplicates of Newton, may all have discovered gravity before *our* Newton did—if you believe that “before” means anything relative to those kinds of separations.

When that thought first occurred to me, I actually found it quite uplifting. Once I realized that someone, somewhere in the expanses of space and time, already knows the answer to any answerable question—even biology questions and history questions; there are other decoherent Earths—then I realized how silly it was to think as if the joy of discovery ought to be limited to one person. It becomes a fully inescapable source of unresolvable existential angst, and I regard that as a reductio.

The consistent solution which maintains the *possibility* of fun, is to stop worrying about what other people know. If you don't know the answer, it's a mystery to you. If you can raise your hand, and clench your fingers into a fist, and you've got no idea of how your brain is doing it—or even what exact muscles lay beneath your skin—you've got to consider yourself just as ignorant as a hunter-gatherer. Sure, someone else knows the answer—but back in the hunter-gatherer days, someone else in an alternate Earth, or for that matter, someone else in the future, knew what the answer was.

Mystery, and the joy of finding out, is either a personal thing, or it doesn't exist at all—and I prefer to say it's personal.

The joy of assisting your civilization by telling it something it doesn't already know, does tend to be one-shot per discovery per civilization; that kind of value is conserved, as are Nobel Prizes. And the prospect of that reward may be what it takes to keep you focused on one problem for the years required to develop a really *deep* understanding; plus, working on a problem unknown to your civilization is a sure-fire way to avoid reading any spoilers.

But as part of my general project to undo this idea that rationalists have less fun, I want to restore the magic and mystery to every part of the world which you do not *personally* understand, regardless of what other knowledge may exist, far away in space and time, or even in your next-door neighbor's mind. If *you* don't know, it's a mystery. And now think of how many things you don't know! (If you can't think of anything, you have other problems<sup>5</sup>.) Isn't the world suddenly a much more mysterious and magical and *interesting* place? As if you'd been transported into an alternate dimension, and had to learn all the rules from scratch?

“A friend once told me that I look at the world as if I've never seen it before. I thought, that's a nice compliment... Wait! I never *have* seen it before! What —did everyone else get a preview?”

—Ran Prieur<sup>6</sup>

---

5. Page 140, ‘Guessing the Teacher’s Password’.

6. <http://www.ranprieur.com/me/100things.html>

## 14. Bind Yourself to Reality<sup>1</sup>

**Followup to:** Joy in the Merely Real<sup>2</sup>

So perhaps you're reading all this, and asking: "Yes, but what does this have to do with reductionism?"

Partially, it's a matter of leaving a line of retreat<sup>4</sup>. It's not easy to take something *important* apart into components, when you're convinced that this removes magic from the world, unweaves the rainbow. I do plan to take certain things apart, on this blog; and I prefer not to create pointless existential anguish.

Partially, it's the crusade against Hollywood Rationality, the concept that understanding the rainbow subtracts its beauty. The rainbow is still beautiful *plus* you get the beauty of physics.

But even more deeply, it's one of these subtle hidden-core-of-rationality<sup>5</sup> things. You know, the sort of thing where I start talking about 'the Way'<sup>6</sup>. It's about *binding yourself to reality*.

In one of Frank Herbert's *Dune* books, IIRC, it is said that a Truthsayer gains their ability to detect lies in others by always speaking truth themselves, so that they form a relationship with the truth whose violation they can feel. It wouldn't work, but I still think it's one of the more beautiful thoughts in fiction. At the very least, to get close to the truth, you have to be willing to press yourself up against reality as tightly as possible, without flinching away, or sneering down.

You can see the bind-yourself-to-reality theme in "Lotteries: A Waste of Hope"<sup>7</sup>. Understanding that lottery tickets have negative expected utility, does not mean that you give up the hope of being rich. It means that you stop wasting that hope on lottery tickets. You put the hope into your job, your school, your startup, your eBay sideline; and if you truly have nothing worth hoping for, then maybe it's time to start looking.

---

1. [http://lesswrong.com/lw/ot/bind\\_yourself\\_to\\_reality/](http://lesswrong.com/lw/ot/bind_yourself_to_reality/)

2. Page 715, 'Joy in the Merely Real'.

3. Page 699, 'Reductionism'.

4. Page 653, 'Leave a Line of Retreat'.

5. <http://yudkowsky.net/virtues/>

6. [http://lesswrong.com/lw/m7/zen\\_and\\_the\\_art\\_of\\_rationality/](http://lesswrong.com/lw/m7/zen_and_the_art_of_rationality/)

7. [http://lesswrong.com/lw/bl/lotteries\\_a\\_waste\\_of\\_hope/](http://lesswrong.com/lw/bl/lotteries_a_waste_of_hope/)

It's not dreams I object to, only *impossible* dreams. The lottery isn't impossible, but it is an un-actionable near-impossibility. It's not that winning the lottery is extremely *difficult*—requires a desperate effort—but that *work* isn't the issue.

I say all this, to exemplify the idea of taking emotional energy that is flowing off to nowhere, and binding it into the realms of reality.

This doesn't mean setting goals that are low enough to be "realistic", i.e., easy and safe and parentally approved. Maybe this is good advice in your personal case, I don't know, but I'm not the one to say it.

What I mean is that you can invest emotional energy in rainbows even if they turn out *not* to be magic. The future is always absurd<sup>8</sup> but it is never *unreal*.

The Hollywood Rationality stereotype is that "rational = emotionless"; the more reasonable you are, the more of your emotions Reason inevitably destroys. In "Feeling Rational"<sup>9</sup> I contrast this against "*That which can be destroyed by the truth should be*" and "*That which the truth nourishes should thrive*". When you have arrived at your best picture of the truth, there is nothing irrational about the emotions you feel as a result of that—the emotions cannot be destroyed by truth, so they must not be irrational.

So instead of *destroying* emotional energies associated with bad explanations for rainbows, as the Hollywood Rationality stereotype would have it, let us *redirect* these emotional energies into reality—bind them to beliefs that are as true as we can make them.

Want to fly? Don't give up on flight. Give up on flying potions and build yourself an airplane.

Remember the theme of "Think Like Reality"<sup>10</sup>, where I talked about how when physics seems counterintuitive, you've got to accept that it's not *physics* that's weird, it's *you*?

What I'm talking about now is like that, only with emotions<sup>11</sup> instead of hypotheses—binding your feelings into the real world. Not the "realistic" everyday world. I would be a howling hypocrite

8. [http://lesswrong.com/lw/j6/why\\_is\\_the\\_future\\_so\\_absurd/](http://lesswrong.com/lw/j6/why_is_the_future_so_absurd/)

9. Page 625, 'Feeling Rational'.

10. [http://lesswrong.com/lw/hs/think\\_like\\_reality/](http://lesswrong.com/lw/hs/think_like_reality/)

11. Page 625, 'Feeling Rational'.

if I told you to shut up and do your homework. I mean the *real* real world, the lawful universe<sup>12</sup>, that includes absurdities<sup>13</sup> like Moon landings and the evolution of human intelligence. Just not any magic, anywhere, ever.

It is a Hollywood Rationality meme that “Science takes the fun out of life.”

Science puts the fun back *into* life.

Rationality directs your emotional energies into the universe, rather than somewhere else.

---

12. [http://lesswrong.com/lw/br/universal\\_law/](http://lesswrong.com/lw/br/universal_law/)

13. [http://lesswrong.com/lw/j6/wby\\_is\\_the\\_future\\_so\\_absurd/](http://lesswrong.com/lw/j6/wby_is_the_future_so_absurd/)

## 15. If You Demand Magic, Magic Won't Help<sup>1</sup>

**Followup to:** Explaining vs. Explaining Away<sup>2</sup>, Joy in the Merely Real<sup>3</sup>

Most witches don't believe in gods. They know that the gods exist, of course. They even deal with them occasionally. But they don't believe in them. They know them too well. It would be like believing in the postman.

—Terry Pratchett, *Witches Abroad*

Once upon a time, I was pondering the philosophy of fantasy stories—

And before anyone chides me for my “failure to understand what fantasy is about”, let me say this: I was raised in an SF&F household. I have been reading fantasy stories since I was five years old. I occasionally try to *write* fantasy stories<sup>4</sup>. And I am *not* the sort of person who tries to write for a genre without pondering its philosophy. Where do you think story ideas come from?

Anyway:

I was pondering the philosophy of fantasy stories, and it occurred to me that if there were actually dragons in our world—if you could go down to the zoo, or even to a distant mountain, and meet a fire-breathing dragon—while nobody had ever actually seen a zebra, then our fantasy stories would contain zebras aplenty, while dragons would be unexciting.

Now that's what I call painting yourself into a corner, wot? The grass is always greener on the other side of unreality.

In one of the standard fantasy plots, a protagonist from our Earth, a sympathetic character with lousy grades or a crushing mortgage but still a good heart, suddenly finds themselves in a world<sup>5</sup> where magic operates in place of science. The protagonist often goes on to practice magic, and become in due course a (super-powerful) sorcerer.

---

1. [http://lesswrong.com/lw/ou/if\\_you\\_demand\\_magic\\_magic\\_wont\\_help/](http://lesswrong.com/lw/ou/if_you_demand_magic_magic_wont_help/)

2. Page 703, ‘Explaining vs. Explaining Away’.

3. Page 715, ‘Joy in the Merely Real’.

4. <http://yudkowsky.net/other/fiction>

5. [http://lesswrong.com/lw/hq/universal\\_fire/](http://lesswrong.com/lw/hq/universal_fire/)

Now here's the question—and yes, it is a little unkind, but I think it needs to be asked: Presumably most readers of these novels see themselves in the protagonist's shoes, fantasizing about their own acquisition of sorcery. Wishing for magic. And, barring improbable demographics, most readers of these novels are not scientists.

Born into a world of science, they did not become scientists. What makes them think that, in a world of magic, they would act any differently?

If they don't have the scientific attitude, that nothing is "mere"<sup>6</sup>—the capacity to be interested in merely real things—how will magic help them? If they actually *had* magic, it would be merely *real*, and lose the charm of unattainability. They might be excited at first, but (like the lottery winners who, six months later, aren't nearly as happy as they expected to be), the excitement would soon wear off. Probably as soon as they had to actually *study* spells.

*Unless* they can find the capacity to take joy in things that are merely real. To be just as excited by hang-gliding, as riding a dragon; to be as excited by making a light with electricity, as by making a light with magic... even if it takes a little study...

Don't get me wrong. I'm not dissing dragons. Who knows, we might even create some, one of these days.

But if you don't have the capacity to enjoy hang-gliding even though it is *merely real*, then as soon as dragons *turn* real, you're not going to be any more excited by dragons than you are by hang-gliding.

Do you think you would prefer living in the Future, to living in the present? That's a quite understandable preference. Things do seem to be getting better over time.

But don't forget that *this is* the Future, relative to the Dark Ages of a thousand years earlier. You have opportunities undreamt-of even by kings.

If the trend continues, the Future might be a very fine place indeed in which to live. But if you do make it to the Future, what you find, when you get there, will be another Now. If you don't have the basic capacity to enjoy being in a Now—if your emotional en-

---

6. Page 715, 'Joy in the Merely Real'.

ergy can *only* go into the Future, if you can *only* hope for a better tomorrow—then no amount of passing time can help you.

(Yes, in the Future there could be a pill that fixes the emotional problem of always looking to the Future. I don't think this invalidates my basic point, which is about what sort of pills we should want to take.)

Matthew C., commenting here on LW<sup>7</sup>, seems very excited about an informally specified “theory” by Rupert Sheldrake which “explains<sup>8</sup>” such non-explanation-demanding phenomena as protein folding and snowflake symmetry. But why isn’t Matthew C. just as excited about, say, Special Relativity? Special Relativity is actually *known* to be a law, so why isn’t it even *more* exciting? The advantage of becoming excited about a law already known to be true, is that you know your excitement will not be wasted.

If Sheldrake’s theory were accepted truth taught in elementary schools, Matthew C. wouldn’t care about it. Or why else is Matthew C. fascinated by that one particular law which he believes to be a law of physics, more than all the other laws?

The worst catastrophe you could visit upon the New Age community would be for their rituals to start working reliably, and for UFOs to actually appear in the skies. What would be the point of believing in aliens, if they were just *there*, and everyone else could see them too? In a world where psychic powers were merely real, New Agers wouldn’t *believe in* psychic powers, any more than anyone cares enough about gravity to believe in it. (Except for scientists, of course.)

Why am I so negative about magic? Would it be *wrong* for magic to exist?

I’m not actually negative on magic. Remember, I occasionally try to write fantasy stories. But I’m annoyed with this psychology that, if it were born into a world where spells and potions did work, would pine away for a world where household goods were abundantly produced by assembly lines.

Part of binding yourself to reality, on an emotional as well as intellectual level, is coming to terms with the fact that you *do live here*.

---

7. <http://lesswrong.com/lw/on/reductionism/irb>

8. Page 147, ‘Fake Causality’.

Only then can you see this, your world, and whatever opportunities it holds out for you, without wishing your sight away.

Not to put too fine a point on it, but *I've* found no lack of dragons to fight, or magics to master, in this world of my birth. If I were transported into one of those fantasy novels, I wouldn't be surprised to find myself studying the forbidden ultimate sorcery—

—because why should being transported into a magical world change anything? It's not *where* you are, it's *who* you are.

So remember the Litany Against Being Transported Into An Alternate Universe:

If I'm going to be happy anywhere,  
Or achieve greatness anywhere,  
Or learn true secrets anywhere,  
Or save the world anywhere,  
Or feel strongly anywhere,  
Or help people anywhere,  
I may as well do it in reality.

## 16. Mundane Magic<sup>1</sup>

**Followup to:** Joy in the Merely Real<sup>2</sup>, Joy in Discovery<sup>3</sup>, If You Demand Magic, Magic Won't Help<sup>4</sup>

As you may recall from some months earlier, I think that part of the rationalist ethos is *binding yourself emotionally* to an absolutely lawful<sup>5</sup> reductionistic<sup>6</sup> universe—a universe containing no ontologically basic mental things<sup>7</sup> such as souls or magic<sup>8</sup>—and pouring all your hope and all your care into that merely real universe and its possibilities, without disappointment.

There's an old trick for combating dukkha<sup>9</sup> where you make a list of things you're grateful for, like a roof over your head.

So why not make a list of abilities you have that would be amazingly cool *if they were magic*, or if only a few chosen individuals had them<sup>10</sup>?

For example, suppose that instead of one eye, you possessed a magical *second* eye embedded in your forehead. And this second eye enabled you to *see into the third dimension*—so that you could somehow tell *how far away* things were—where an ordinary eye would see only a two-dimensional shadow of the true world. Only the possessors of this ability can accurately aim the legendary distance-weapons that kill at ranges far beyond a sword, or use to their fullest potential the shells of ultrafast machinery called “cars”.

“Binocular vision” would be too light a term<sup>11</sup> for this ability. We'll only appreciate it once it has a properly impressive name, like Mystic Eyes of Depth Perception.

So here's a list of some of my favorite magical powers:

- *Vibratory Telepathy*. By transmitting invisible vibrations through the very air itself, two users of this ability can

---

1. [http://lesswrong.com/lw/ve/mundane\\_magic/](http://lesswrong.com/lw/ve/mundane_magic/)

2. Page 715, ‘Joy in the Merely Real’.

3. Page 717, ‘Joy in Discovery’.

4. Page 724, ‘If You Demand Magic, Magic Won't Help’.

5. [http://lesswrong.com/lw/hr/universal\\_law/](http://lesswrong.com/lw/hr/universal_law/)

6. Page 699, ‘Reductionism’.

7. Page 826, ‘Excluding the Supernatural’.

8. Page 724, ‘If You Demand Magic, Magic Won't Help’.

9. <http://en.wikipedia.org/wiki/Dukkha>

10. Page 717, ‘Joy in Discovery’.

11. <http://chaitanya1.files.wordpress.com/2008/06/bigbang.gif>

*share thoughts.* As a result, Vibratory Telepaths can form emotional bonds much deeper than those possible to other primates.

- *Psychometric Tracery.* By tracing small fine lines on a surface, the Psychometric Tracer can leave impressions of emotions, history, knowledge, even the structure of other spells. This is a higher level than Vibratory Telepathy as a Psychometric Tracer can share the thoughts of long-dead Tracers who lived thousands of years earlier. By reading one Tracery and inscribing another simultaneously, Tracers can duplicate Tracings; and these replicated Tracings can even contain the detailed pattern of other spells and magics. Thus, the Tracers wield almost unimaginable power as magicians; but Tracers can get in trouble trying to use complicated Traceries that they could not have Traced themselves.
- *Multidimensional Kinesis.* With simple, almost unthinking acts of will, the Kinetics can cause extraordinarily complex forces to flow through small tentacles and into any physical object within touching range—not just pushes, but combinations of pushes at many points that can effectively apply torques and twists. The Kinetic ability is far subtler than it first appears: they use it not only to wield existing objects with martial precision, but also to apply forces that sculpt objects into forms more suitable for Kinetic wielding. They even create tools that extend the power of their Kinesis and enable them to sculpt ever-finer and ever-more-complicated tools, a positive feedback loop fully as impressive as it sounds.
- *The Eye.* The user of this ability can perceive infinitesimal traveling twists in the Force that binds matter—tiny vibrations, akin to the life-giving power of the Sun that falls on leaves, but far more subtle. A bearer of the Eye can sense objects far beyond the range of touch using the tiny disturbances they make in the Force. Mountains many days travel away can be known to them as if within arm's reach. According to the bearers of the Eye, when night falls and sunlight fails, they can sense huge fusion fires burning at unthinkable distances—though no one

else has any way of verifying this. Possession of a single Eye is said to make the bearer equivalent to royalty.  
And finally,

- *The Ultimate Power.* The user of this ability contains a smaller, imperfect echo of the entire universe, enabling them to search out paths through probability to any desired future. If this sounds like a ridiculously powerful ability, you're right—game balance goes right out the window with this one. Extremely rare among life forms, it is the *sekai no ougi* or “hidden technique of the world”.

Nothing can oppose the Ultimate Power except the Ultimate Power. Any less-than-ultimate Power will simply be “comprehended” by the Ultimate and disrupted in some inconceivable fashion, or even absorbed into the Ultimates’ own power base. For this reason the Ultimate Power is sometimes called the “master technique of techniques” or the “trump card that trumps all other trumps”. The more powerful Ultimates can stretch their “comprehension” across galactic distances and aeons of time, and even perceive the bizarre laws of the hidden “world beneath the world”.

Ultimates have been killed by immense natural catastrophes, or by extremely swift surprise attacks that give them no chance to use their power. But all such victories are ultimately a matter of luck—it does not confront the Ultimates on their own probability-bending level, and if they survive they will begin to bend Time to avoid future attacks.

But the Ultimate Power itself is also dangerous, and many Ultimates have been destroyed by their own powers—falling into one of the flaws in their imperfect inner echo of the world.

Stripped of weapons and armor and locked in a cell, an Ultimate is still one of the most dangerous life-forms on the planet. A sword can be broken and a limb can be cut

off, but the Ultimate Power is “the power that cannot be removed without removing you”.

Perhaps because this connection is so intimate, the Ultimates regard one who loses their Ultimate Power permanently—without hope of regaining it—as *sciaovo*, or “dead while breathing”. The Ultimates argue that the Ultimate Power is so important as to be a necessary part of what makes a creature an end in itself, rather than a means. The Ultimates even insist that anyone who lacks the Ultimate Power cannot begin to truly comprehend the Ultimate Power, and hence, cannot understand why the Ultimate Power is morally important—a suspiciously self-serving argument.

The users of this ability form an absolute aristocracy and treat all other life forms as their pawns.

## 17. The Beauty of Settled Science<sup>1</sup>

Facts do not need<sup>2</sup> to be unexplainable, to be beautiful; truths do not become less worth learning<sup>3</sup>, if someone else knows them; beliefs do not become less worthwhile<sup>4</sup>, if many others share them...

...and if you only care about scientific issues that are controversial, you will end up with a head stuffed full of garbage.

The media thinks that only the cutting edge of science is worth reporting on. How often do you see headlines like “General Relativity still governing planetary orbits” or “Phlogiston theory remains false”? So, by the time anything is solid science, it is no longer a breaking headline. “Newsworthy” science is often based on the thinnest of evidence and wrong half the time—if it were not on the uttermost fringes of the scientific frontier, it would not be breaking news.

Scientific *controversies* are problems *so difficult* that even people who’ve spent years mastering the field can still fool themselves. That’s what makes for the heated arguments that attract all the media attention.

Worse, if you aren’t in the field and part of the game, controversies *aren’t even fun*.

Oh, sure, you can have the fun of picking a side in an argument. But you can get that in any football game<sup>5</sup>. That’s not what the fun of science is about.

Reading a well-written textbook, you get: Carefully phrased explanations for incoming students, math derived step by step (where applicable), plenty of experiments cited as illustration (where applicable), test problems on which to display your new mastery, and a reasonably good guarantee that what you’re learning is actually true.

Reading press releases, you usually get: Fake explanations<sup>6</sup> that convey nothing except the delusion of understanding<sup>7</sup> of a result

---

1. [http://lesswrong.com/lw/ow/the\\_beauty\\_of\\_settled\\_science/](http://lesswrong.com/lw/ow/the_beauty_of_settled_science/)

2. Page 715, ‘Joy in the Merely Real’.

3. Page 717, ‘Joy in Discovery’.

4. Page 724, ‘If You Demand Magic, Magic Won’t Help’.

5. [http://lesswrong.com/lw/mg/the\\_twoparty\\_swindle/](http://lesswrong.com/lw/mg/the_twoparty_swindle/)

6. Page 138, ‘Fake Explanations’.

7. Page 140, ‘Guessing the Teacher’s Password’.

that the press release author didn't understand and that probably has a better-than-even chance of failing to replicate.

Modern science is built on discoveries, built on discoveries, built on discoveries, and so on, all the way back to people like Archimedes, who discovered facts like why boats float, that can make sense even if you don't know about other discoveries. A good place to start traveling that road is at the beginning.

Don't be embarrassed to read *elementary* science textbooks, either. If you want to pretend to be sophisticated, go find a play to sneer at. If you just want to have *fun*, remember that simplicity is at the core of scientific beauty.

And thinking you can jump right into the frontier, when you haven't learned the settled science, is like...

...like trying to climb only the *top* half of Mount Everest (which is the only part that interests you) by standing at the base of the mountain, bending your knees, and jumping *really hard* (so you can pass over the boring parts).

Now I'm not saying that you should never pay attention to scientific controversies. If 40% of oncologists think that white socks cause cancer, and the other 60% violently disagree, this is an important fact to know.

Just don't go thinking that science *has* to be controversial to be interesting.

Or, for that matter, that science has to be recent to be interesting. A steady diet of science *news* is bad for you: You are what you eat, and if you eat only science reporting on fluid situations, without a solid textbook now and then, your brain will turn to liquid.

## 18. Amazing Breakthrough Day: April 1st<sup>1</sup>

So you're thinking, "April 1st... isn't that already supposed to be April Fool's Day?"

Yes—and that will provide the ideal cover for celebrating Amazing Breakthrough Day.

As I argued in "The Beauty of Settled Science"<sup>2</sup>, it is a major problem that media coverage of science focuses only on *breaking news*. Breaking news, in science, occurs at the furthest fringes of the scientific frontier, which means that the new discovery is often:

- Controversial
- Supported by only one experiment
- Way the heck more complicated than an ordinary mortal can handle, and requiring lots of prerequisite science to understand, which is why it wasn't solved three centuries ago
- Later shown to be wrong

People never get to see the *solid* stuff, let alone the *understandable* stuff, because it isn't *breaking news*.

On Amazing Breakthrough Day, I propose, journalists who really care about science can report—under the protective cover of April 1st—such important but neglected science stories as:

- BOATS EXPLAINED: Centuries-Old Problem Solved By Bathtub Nudist<sup>3</sup>
- YOU SHALL NOT CROSS! Königsberg<sup>4</sup> Tourists' Hopes Dashed
- ARE YOUR LUNGS ON FIRE? Link Between Respiration And Combustion<sup>5</sup> Gains Acceptance Among Scientists

Note that every one of these headlines are *true*—they describe events that did, in fact, happen. They just didn't happen *yesterday*.

There have been many humanly understandable amazing breakthroughs in the history of science, which can be understood with-

---

1. [http://lesswrong.com/lw/ox/amazing\\_breakthrough\\_day\\_april\\_1st/](http://lesswrong.com/lw/ox/amazing_breakthrough_day_april_1st/)

2. Page 732, 'The Beauty of Settled Science'.

3. <http://en.wikipedia.org/wiki/Archimedes>

4. [http://en.wikipedia.org/wiki/Seven\\_Bridges\\_of\\_K%C3%B6nigsberg](http://en.wikipedia.org/wiki/Seven_Bridges_of_K%C3%B6nigsberg)

5. [http://lesswrong.com/lw/hq/universal\\_fire/](http://lesswrong.com/lw/hq/universal_fire/)

out a PhD or even BSc. The operative word here is *history*. Think of Archimedes’s “Eureka!” when he understood the relation between the water a ship displaces, and the reason the ship floats. This is *far enough back* in scientific history that you don’t need to know 50 other discoveries to understand the theory; it can be explained in a couple of graphs; anyone can see how it’s useful; and the confirming experiments can be duplicated in your own bathtub.

Modern science is built on discoveries built on discoveries built on discoveries and so on all the way back to Archimedes. Reporting science *only* as breaking news is like wandering into a movie 3/4ths of the way through, writing a story about “Bloody-handed man kisses girl holding gun!” and wandering back out again.

And if your editor says, “Oh, but our readers won’t be interested in that—”

Then point out that Reddit and Digg don’t link *only* to breaking news. They also link to short webpages that give good explanations of old science. Readers vote it up, and that should tell you something. Explain that if your newspaper doesn’t change to look more like Reddit, you’ll have to start selling drugs to make payroll. Editors love to hear that sort of thing, right?

On the Internet, a good new explanation of old science *is* news and it spreads like news. Why couldn’t the science sections of newspapers work the same way? Why isn’t a new *explanation* worth reporting on?

But all this is too visionary for a first step. For now, let’s just see if any journalists out there pick up on Amazing Breakthrough Day, where you report on some *understandable* science breakthrough as though it had just occurred.

April 1st. Put it on your calendar.

## 19. Is Humanism A Religion-Substitute?<sup>1</sup>

### Followup to: Bind Yourself to Reality<sup>2</sup>

For many years before the Wright Brothers, people dreamed of flying with magic potions. There was nothing irrational<sup>3</sup> about the *raw desire* to fly. There was nothing *tainted* about the wish to look down on a cloud from above. Only the “magic potions” part was irrational.

Suppose you were to put me into an fMRI scanner, and take a movie of my brain’s activity levels, while I watched a space shuttle launch. (Wanting to visit space is not “realistic”, but it is an essentially lawful dream—one that can be fulfilled in a lawful universe.) The fMRI might—maybe, maybe not—resemble the fMRI of a devout Christian watching a nativity scene.

Should an experimenter obtain this result, there’s a lot of people out there, both Christians and some atheists, who would gloat: “Ha, ha, space travel is your religion!<sup>4</sup>!”

But that’s drawing the wrong category<sup>5</sup> boundary<sup>6</sup>. It’s like saying that, because some people once tried to fly by irrational means, no one should ever enjoy looking out of an airplane window on the clouds below.

If a rocket launch is what it takes to give me a feeling of aesthetic transcendence, I do not see this as a *substitute* for religion. That is theomorphism—the viewpoint from gloating religionists who assume that everyone who *isn’t* religious has a hole in their mind that wants filling.

Now, to be fair to the religionists, this is not *just* a gloating assumption. There *are* atheists who have religion-shaped holes in their minds. I *have* seen attempts to substitute atheism or even transhumanism for religion. And the result is invariably awful. Utterly awful. Absolutely abjectly awful.

I call such efforts, “hymns to the nonexistence of God”.

---

1. [http://lesswrong.com/lw/oy/is\\_humanism\\_a\\_religionsubstitute/](http://lesswrong.com/lw/oy/is_humanism_a_religionsubstitute/)

2. Page 721, ‘Bind Yourself to Reality’.

3. Page 721, ‘Bind Yourself to Reality’.

4. Page 392, ‘Uncritical Supercriticality’.

5. Page 275, ‘Mutual Information, and Density in Thingspace’.

6. Page 282, ‘Superexponential Conceptspace, and Simple Words’.

When someone sets out to write an atheistic hymn—"Hail, oh unintelligent universe," blah, blah, blah—the result will, without exception, suck.

Why? Because they're being imitative.<sup>7</sup> Because they have no motivation for writing the hymn *except* a vague feeling that since churches have hymns, they ought to have one too. And, on a purely artistic level, that puts them far beneath genuine religious art that is not an imitation of anything, but an original expression of emotion.

Religious hymns were (often) written by people who *felt strongly* and *wrote honestly* and put serious effort into the prosody and imagery of their work—that's what gives their work the grace that it possesses, of artistic integrity.

So are atheists doomed to hymnlessness?

There is an acid test of attempts at post-theism. The acid test is: "If religion had never existed among the human species—if we had *never made* the original mistake—would this song, this art, this ritual, this way of thinking, still make sense?"

If humanity had never made the original mistake, there would be no hymns to the nonexistence of God. But there would still be marriages, so the notion of an atheistic marriage ceremony makes perfect sense—as long as you don't suddenly launch into a lecture on how God doesn't exist. Because, in a world where religion *never had* existed, nobody would interrupt a wedding to talk about the implausibility of a distant hypothetical concept. They'd talk about love, children, commitment, honesty, devotion, but who the heck would mention God?

And, in a human world where religion *never had* existed, there would still be people who got tears in their eyes watching a space shuttle launch.

Which is why, even if experiment shows that watching a shuttle launch makes "religion"-associated areas of my brain light up, associated with feelings of transcendence, I do not see that as a *substitute* for religion; I expect the same brain areas would light up, for the same reason, if I lived in a world where religion had never been invented.

A good "atheistic hymn" is simply a song about anything worth singing about that doesn't happen to be religious.

---

<sup>7</sup>. Page 451, 'Cached Thoughts'.

Also, reversed stupidity is not intelligence<sup>8</sup>. The world's greatest idiot may say the Sun is shining, but that doesn't make it dark out. The point is *not* to create a life that resembles religion as little as possible in every surface aspect—this is the same kind of thinking that inspires hymns to the nonexistence of God. If humanity had never made the original mistake, no one would be *trying to avoid* things that vaguely resembled religion. Believe accurately, then feel accordingly<sup>9</sup>: If space launches actually exist, and watching a rocket rise makes you want to sing, then write the song, dammit.

If I get tears in my eyes at a space shuttle launch, it doesn't mean I'm trying to fill a hole left by religion—it means that my emotional energies, my *caring*, are bound into the real world<sup>10</sup>.

If God did speak plainly, and answer prayers reliably, God would just become one more boringly real thing, no more worth believing in than the postman<sup>11</sup>. If God were real, it would destroy the inner uncertainty that brings forth outward fervor in compensation. And if everyone else believed God were real, it would destroy the specialness of being one of the elect.

If you invest your emotional energy in space travel, you don't have those vulnerabilities. I can *see* the Space Shuttle rise without losing the awe. Everyone else can believe that Space Shuttles are real, and it doesn't make them any less special. I haven't painted myself into the corner.

The choice between God and humanity is not just a choice of drugs. Above all, humanity *actually exists*.

---

8. Page 335, 'Reversed Stupidity Is Not Intelligence'.

9. Page 625, 'Feeling Rational'.

10. Page 721, 'Bind Yourself to Reality'.

11. Page 724, 'If You Demand Magic, Magic Won't Help'.

## 20. Scarcity<sup>1</sup>

What follows is taken primarily from Robert Cialdini's *Influence: The Psychology of Persuasion*. I own three copies of this book, one for myself, and two for loaning to friends.

*Scarcity*, as that term is used in social psychology, is when things become *more desirable* as they appear *less obtainable*.

- If you put a two-year-old boy in a room with two toys, one toy in the open and the other behind a Plexiglas wall, the two-year-old will ignore the easily accessible toy and go after the apparently forbidden one. If the wall is low enough to be easily climbable, the toddler is no more likely to go after one toy than the other. (Brehm and Weintraub 1977.)
- When Dade County forbade use or possession of phosphate detergents, many Dade residents drove to nearby counties and bought huge amounts of phosphate laundry detergents. Compared to Tampa residents not affected by the regulation, Dade residents rated phosphate detergents as gentler, more effective, more powerful on stains, and even believed that phosphate detergents poured more easily. (Mazis 1975, Mazis et. al. 1973.)

Similarly, information that appears forbidden or secret, seems more important and trustworthy:

- When University of North Carolina students learned that a speech opposing coed dorms had been banned, they became more opposed to coed dorms (without even hearing the speech). (Probably in Ashmore et. al. 1971.)
- When a driver said he had liability insurance, experimental jurors awarded his victim an average of four thousand dollars more than if the driver said he had no insurance. If the judge afterward informed the jurors that information about insurance was inadmissible and must be ignored, jurors awarded an average of thirteen thousand dollars more than if the driver had no insurance. (Broeder 1959.)

---

<sup>1</sup>. <http://lesswrong.com/lw/oz/scarcity/>

- Buyers for supermarkets, told by a supplier that beef was in scarce supply, gave orders for twice as much beef as buyers told it was readily available. Buyers told that beef was in scarce supply, and furthermore, that the information about scarcity was itself scarce—that the shortage was not general knowledge—ordered six times as much beef. (Since the study was conducted in a real-world context, the information provided was in fact correct.) (Knishinsky 1982.)

The conventional theory for explaining this is “psychological reactance”, social-psychology-speak for “When you tell people they can’t do something, they’ll just try even harder.” The fundamental instincts involved appear to be preservation of status and preservation of options. We resist dominance, when any human agency tries to restrict our freedom. And when options seem to be in danger of disappearing, even from natural causes, we try to leap on the option before it’s gone.

Leaping on disappearing options may be a good adaptation in a hunter-gatherer<sup>2</sup> society—gather the fruits while the tree is still in bloom—but in a money-based society it can be rather costly. Cialdini (1993) reports that in one appliance store he observed, a salesperson who saw that a customer was evincing signs of interest in an appliance would approach, and sadly inform the customer that the item was out of stock, the last one having been sold only twenty minutes ago. Scarcity creating a sudden jump in desirability, the customer would often ask whether there was any chance that the salesperson could locate an unsold item in the back room, warehouse, or anywhere. “Well,” says the salesperson, “that’s possible, and I’m willing to check; but do I understand that this is the model you want, and if I can find it at this price, you’ll take it?”

As Cialdini remarks, a chief sign of this malfunction is that you dream of *possessing* something, rather than *using* it. (Timothy Ferriss offers similar advice on planning your life: ask which *ongoing experiences* would make you happy, rather than which possessions or status-changes.)

But the really fundamental problem with desiring the unattainable is that as soon as you actually *get* it, it stops being unattain-

---

<sup>2.</sup> [http://lesswrong.com/lw/lo/adaptationexecuters\\_not\\_fitnessmaximizers/](http://lesswrong.com/lw/lo/adaptationexecuters_not_fitnessmaximizers/)

able<sup>3</sup>. If we cannot take joy in the merely available, our lives will always be frustrated...

---

Ashmore, R. D., Ramachandra, V. and Jones, R. A. (1971.) "Censorship as an Attitude Change Induction." Paper presented at Eastern Psychological Association meeting, New York, April 1971.

Brehm, S. S. and Weintraub, M. (1977.) "Physical Barriers and Psychological Reactance: Two-year-olds' Responses to Threats to Freedom." *Journal of Personality and Social Psychology*, **35**: 830-36.

Broeder, D. (1959.) "The University of Chicago Jury Project." *Nebraska Law Review* **38**: 760-74.

Cialdini, R. B. (1993.) *Influence: The Psychology of Persuasion: Revised Edition*. Pp. 237-71. New York: Quill.

Knishinsky, A. (1982.) "The Effects of Scarcity of Material and Exclusivity of Information on Industrial Buyer Perceived Risk in Provoking a Purchase Decision." Doctoral dissertation, Arizona State University.

Mazis, M. B. (1975.) "Antipollution Measures and Psychological Reactance Theory: A Field Experiment." *Journal of Personality and Social Psychology* **31**: 654-66.

Mazis, M. B., Settle, R. B. and Leslie, D. C. (1973.) "Elimination of Phosphate Detergents and Psychological Reactance." *Journal of Marketing Research* **10**: 390-95.

---

3. Page 724, 'If You Demand Magic, Magic Won't Help'.

## 21. To Spread Science, Keep It Secret<sup>1</sup>

**Followup to:** Joy in Discovery<sup>2</sup>, Bind Yourself to Reality<sup>3</sup>, Scientific Evidence<sup>4</sup>, Scarcity<sup>5</sup>

Sometimes I wonder if the Pythagoreans had the right idea.

Yes, I've written<sup>6</sup> about how "science" is inherently public. I've written that "science" is distinguished from merely rational knowledge by the in-principle ability to reproduce scientific experiments for yourself, to know without relying on authority. I've said that "science" should be defined as the publicly accessible knowledge of humankind. I've even suggested that future generations will regard all papers not published in an open-access journal as non-science, i.e., it can't be part of the public knowledge of humankind if you make people pay to read it.

But that's only one vision of the future. In another vision, the knowledge we now call "science" is taken *out* of the public domain—the books and journals hidden away, guarded by mystic cults<sup>7</sup> of gurus<sup>8</sup> wearing robes<sup>9</sup>, requiring fearsome initiation rituals for access—so that more people will *actually* study it.

I mean, right now, people *can* study science but they *don't*.

"Scarcity"<sup>10</sup>, it's called in social psychology<sup>11</sup>. What appears to be in limited supply, is more highly valued. And this effect is *especially* strong with information—we're much more likely to try to obtain information that we believe is secret, and to value it more when we do obtain it.

With science, I think, people assume that if the information is freely available, it must not be important. So instead people join cults that have the sense to keep their Great Truths secret. The

---

1. [http://lesswrong.com/lw/po/to\\_spread\\_science\\_keep\\_it\\_secret/](http://lesswrong.com/lw/po/to_spread_science_keep_it_secret/)

2. Page 717, 'Joy in Discovery'.

3. Page 721, 'Bind Yourself to Reality'.

4. [http://lesswrong.com/lw/in/scientific\\_evidence\\_legal\\_evidence\\_rational/](http://lesswrong.com/lw/in/scientific_evidence_legal_evidence_rational/)

5. Page 739, 'Scarcity'.

6. [http://lesswrong.com/lw/in/scientific\\_evidence\\_legal\\_evidence\\_rational/](http://lesswrong.com/lw/in/scientific_evidence_legal_evidence_rational/)

7. Page 432, 'Cultish Countercultishness'.

8. Page 349, 'The Litany Against Gurus'.

9. Page 422, 'Two Cult Koans'.

10. Page 739, 'Scarcity'.

11. <http://www.google.com/search?q=scarcity+psychology>

Great Truth may actually be gibberish, but it's more satisfying than coherent science, because it's *secret*.

Science is the great Purloined Letter of our times, left out in the open and ignored.

Sure, scientific openness helps the scientific elite. They've already *been* through the initiation rituals. But for the rest of the planet, science is kept secret a hundred times more effectively by making it freely available, than if its books were guarded in vaults and you had to walk over hot coals to get access. (This being a fearsome trial indeed, since the great secrets of insulation are only available to Physicist-Initiates of the Third Level.)

If scientific knowledge were hidden in ancient vaults (rather than hidden in inconvenient pay-for-access journals), at least then people would *try* to get into the vaults. They'd be *desperate* to learn science. Especially when they saw the power that Eighth Level Physicists could wield, and were told that they *weren't allowed to know* the explanation.

And if you tried to start a cult around oh, say, Scientology, you'd get some degree of public interest, at first. But people would very quickly start asking uncomfortable questions like "Why haven't you given a public demonstration of your Eighth Level powers, like the Physicists?" and "How come none of the Master Mathematicians seem to want to join your cult?" and "Why should I follow your Founder when he isn't an Eighth Level anything outside his own cult?" and "Why should I study *your* cult *first*, when the Dentists of Doom can do things that are so much more impressive?"

When you look at it from that perspective, the escape of math from the Pythagorean cult starts to look like a major strategic blunder for humanity.

Now, I know what you're going to say: "But science *is* surrounded by fearsome initiation rituals! Plus it's *inherently* difficult to learn! Why doesn't *that* count?" Because the public *thinks* that science is freely available, that's why. If you're *allowed* to learn, it must not be important enough *to* learn.

It's an image problem, people taking their cues from others' attitudes. Just *anyone* can walk into the supermarket and buy a light bulb, and nobody looks at it with awe and reverence. The physics supposedly aren't secret (even though *you* don't know<sup>12</sup>), and there's

a one-paragraph explanation<sup>13</sup> in the newspaper that sounds vaguely authoritative and convincing—essentially, no one treats the lightbulb as a sacred mystery, so neither do you.

Even the simplest little things, completely inert objects like crucifixes, can become magical if everyone *looks* at them like they're magic. But since you're theoretically *allowed* to know why the light bulb works without climbing the mountain to find the remote Monastery of Electricians, there's no need to *actually* bother to learn.

Now, because science does in fact have initiation rituals both social and cognitive, scientists are not wholly dissatisfied with their science. The problem is that, in the present world, very few people bother to study science in the first place. Science cannot be the true Secret Knowledge, because just anyone is allowed to know it—*even though, in fact, they don't*.

If the Great Secret of Natural Selection, passed down from Darwin Who Is Not Forgotten, was only ever imparted to you after you paid \$2000 and went through a ceremony involving torches and robes and masks and sacrificing an ox, *then* when you were shown the fossils, and shown the optic cable going through the retina<sup>14</sup> under a microscope, and finally told the Truth, you would say “That's the most brilliant thing ever!” and *be satisfied*. After that, if some other cult tried to tell you it was actually a bearded man in the sky 6000 years ago, you'd laugh like hell.

And you know, it might actually be more *fun* to do things that way. Especially if the initiation required you to put together some of the evidence for yourself—together, or with classmates—before you could tell your Science Sensei you were ready to advance to the next level. It wouldn't be *efficient*, sure, but it would be *fun*.

If humanity had never made the mistake—never gone down the religious path, and never learned to fear anything that smacks of religion—then maybe the Ph.D. granting ceremony would involve litanies and chanting, because, hey, that's what people like. Why take the fun out of everything?

Maybe we're just doing it wrong.

12. Page 717, ‘Joy in Discovery’.

13. Page 138, ‘Fake Explanations’.

14. [http://lesswrong.com/lw/kr/an\\_alien\\_god/](http://lesswrong.com/lw/kr/an_alien_god/)

And no, I'm not *seriously* proposing that we try to reverse the last five hundred years of openness and classify all the science secret. At least, not at the moment. Efficiency is important for now, especially in things like medical research. I'm just explaining why it is that I won't tell anyone the Secret of how the ineffable difference between blueness and redness arises from mere atoms<sup>15</sup> for less than \$100,000—

Ahem! I meant to say, I'm telling you about this vision of an alternate Earth, so that you give science equal treatment with cults. So that you don't undervalue scientific truth when you learn it, *just* because it doesn't seem to be protected appropriately to its value. *Imagine* the robes and masks. Visualize yourself creeping into the vaults and stealing the Lost Knowledge of Newton. And don't be fooled by any organization that *does* use robes and masks, unless they also show you the data.

People seem to have holes in their minds<sup>16</sup> for Esoteric Knowledge, Deep Secrets, the Hidden Truth. And I'm not even criticizing this psychology! There *are* deep secret esoteric hidden truths, like quantum mechanics or Bayes-structure<sup>17</sup>. We've just gotten into the habit of presenting the Hidden Truth in a very *unsatisfying* way, wrapped up in false mundanity.

But if the holes for secret knowledge are not filled by true beliefs, they will be filled by false beliefs. There is *nothing but* science to learn—the emotional energy must either be invested in reality<sup>18</sup>, or wasted in total nonsense, or destroyed. For myself, I think it is better to invest the emotional energy; fun should not be needlessly cast away.

Right now, we've got the worst of both worlds. Science isn't *really* free, because the courses are expensive and the textbooks are expensive. But the public *thinks* that anyone is allowed to know, so it must not be important.

Ideally, you would want to arrange things the other way around.

---

15. Page 679, 'Wrong Questions'.

16. Page 736, 'Is Humanism A Religion-Substitute?'.

17. [http://lesswrong.com/lw/07/searching\\_for\\_bayesstructure/](http://lesswrong.com/lw/07/searching_for_bayesstructure/)

18. Page 721, 'Bind Yourself to Reality'.

## 22. Initiation Ceremony<sup>1</sup>

The torches that lit the narrow stairwell burned intensely and in the wrong color, flame like melting gold or shattered suns.

192... 193...

Brennan's sandals clicked softly on the stone steps, snicking in sequence, like dominos very slowly falling.

227... 228...

Half a circle ahead of him, a trailing fringe of dark cloth whispered down the stairs, the robed figure itself staying just out of sight.

239... 240...

*Not much longer,* Brennan predicted to himself, and his guess was accurate:

Sixteen times sixteen steps was the number, and they stood before the portal of glass.

The great curved gate had been wrought with cunning, humor, and close attention to indices of refraction: it warped light, bent it, folded it, and generally abused it, so that there were hints of what was on the other side (stronger light sources, dark walls) but no possible way of *seeing through*—unless, of course, you had the key: the counter-door, thick for thin and thin for thick, in which case the two would cancel out.

From the robed figure beside Brennan, two hands emerged, gloved in reflective cloth to conceal skin's color. Fingers like slim mirrors grasped the handles of the warped gate—handles that Brennan had not guessed; in all that distortion, shapes could only be anticipated, not seen.

“Do you want to know?” whispered the guide; a whisper nearly as loud as an ordinary voice, but not revealing the slightest hint of gender.

Brennan paused. The answer to the question seemed suspiciously, indeed extraordinarily obvious, even for ritual.

“Yes,” Brennan said finally.

The guide only regarded him silently.

“Yes, I want to know,” said Brennan.

“Know *what*, exactly?” whispered the figure.

Brennan's face scrunched up in concentration, trying to visualize

---

<sup>1</sup>. [http://lesswrong.com/lw/p1/initiation\\_ceremony/](http://lesswrong.com/lw/p1/initiation_ceremony/)

the game to its end, and hoping he hadn't blown it already; until finally he fell back on the first and last resort, which is the truth:

"It doesn't matter," said Brennan, "the answer is still yes."

The glass gate parted down the middle, and slid, with only the tiniest scraping sound, into the surrounding stone.

The revealed room was lined, wall-to-wall, with figures robed and hooded in light-absorbing cloth. The straight walls were not themselves black stone, but mirrored, tiling a square grid of dark robes out to infinity in all directions; so that it seemed as if the people of some much vaster city, or perhaps the whole human kind, watched in assembly. There was a hint of moist warmth in the air of the room, the breath of the gathered: a scent of crowds.

Brennan's guide moved to the center of the square, where burned four torches of that relentless yellow flame. Brennan followed, and when he stopped, he realized with a slight shock that all the cowled hoods were now looking directly at him. Brennan had never before in his life been the focus of such absolute attention; it was frightening, but not entirely unpleasant.

"He is here," said the guide in that strange loud whisper.

The endless grid of robed figures replied in one voice: perfectly blended, exactly synchronized, so that not a single individual could be singled out from the rest, and betrayed:

*"Who is absent?"*

"Jakob Bernoulli," intoned the guide, and the walls replied:

*"Is dead but not forgotten."*

*"Abraham de Moivre,"*

*"Is dead but not forgotten."*

*"Pierre-Simon Laplace,"*

*"Is dead but not forgotten."*

*"Edwin Thompson Jaynes,"*

*"Is dead but not forgotten."*

"They died," said the guide, "and they are lost to us; but we still have each other, and the project continues."

In the silence, the guide turned to Brennan, and stretched forth a hand, on which rested a small ring of nearly transparent material.

Brennan stepped forward to take the ring—

But the hand clenched tightly shut.

"If three-fourths of the humans in this room are women," said the guide, "and three-fourths of the women and half of the men belong

to the Heresy of Virtue, and I am a Virtuist, what is the probability that I am a man?"

"Two-elevenths," Brennan said confidently.

There was a moment of absolute silence.

Then a titter of shocked laughter.

The guide's whisper came again, truly quiet this time, almost nonexistent: "It's one-sixth, actually."

Brennan's cheeks were flaming so hard that he thought his face might melt off. The instinct was very strong to run out of the room and up the stairs and flee the city and change his name and start his life over again and get it right this time.

"An honest mistake is at least honest," said the guide, louder now, "and we may know the honesty by its relinquishment. If I am a Virtuist, what is the probability that I am a man?"

"One—" Brennan started to say.

Then he stopped. Again, the horrible silence.

"Just say 'one-sixth' already," stage-whispered the figure, this time loud enough for the walls to hear; then there was more laughter, not all of it kind.

Brennan was breathing rapidly and there was sweat on his forehead. If he was wrong about this, he really *was* going to flee the city. "Three fourths women times three fourths Virtuists is nine sixteenths female Virtuists in this room. One fourth men times one half Virtuists is two sixteenths male Virtuists. If I have only that information and the fact that you are a Virtuist, I would then estimate odds of two to nine, or a probability of two-elevenths, that you are male. Though I do not, in fact, believe the information given is correct. For one thing, it seems too neat. For another, there are an odd number of people in this room."

The hand stretched out again, and opened.

Brennan took the ring. It looked almost invisible, in the torch-light; not glass, but some material with a refractive index very close to air. The ring was warm from the guide's hand, and felt like a tiny living thing as it embraced his finger.

The relief was so great that he nearly didn't hear the cowled figures applauding.

From the robed guide came one last whisper:

"You are now a novice of the Bayesian Conspiracy."

Elimonk2darker

Image: *The Bayesian Master*, by Erin Devereux

## 23. Awww, a Zebra<sup>1</sup>

This image recently showed up on Flickr (original is nicer):

Zebra\_4

<sup>2</sup>

With the caption:

*“Alas for those who turn their eyes from zebras and dream of dragons! If we cannot learn to take joy in the merely real, our lives shall be empty indeed.” —Eliezer S. Yudkowsky.*

“Awww!”, I said, and called over my girlfriend over to look.

“Awww!”, she said, and then looked at me, and said, “I think you need to take your own advice!”

Me: “But I’m looking at the zebra!”

Her: “On a computer?”

Me: (*Turns away, hides face.*)

Her: “Have you ever even *seen* a zebra in real life?”

Me: “Yes! Yes, I have! My parents took me to Lincoln Park Zoo! ...man, I hated that place.”

---

1. [http://lesswrong.com/lw/uf/awww\\_a\\_zebra/](http://lesswrong.com/lw/uf/awww_a_zebra/)

2. <http://www.flickr.com/photos/bostian/2901943039/>

## 24. Hand vs. Fingers<sup>1</sup>

**Followup to:** Reductionism<sup>2</sup>, Explaining vs. Explaining Away<sup>3</sup>, Fake Reductionism<sup>4</sup>

Back to our original topic: Reductionism<sup>5</sup>, which (in case you've forgotten) is part of a sequence on the Mind Projection Fallacy<sup>6</sup>. There can be emotional problems in accepting reductionism, if you think that things have to be fundamental to be fun. But this position commits us to never taking joy<sup>7</sup> in anything more complicated than a quark, and so I prefer to reject it.

To review, the reductionist thesis is that we use multi-level models for computational reasons, but physical reality has only a single level. If this doesn't sound familiar, please reread "Reductionism"<sup>8</sup>.

---

Today I'd like to pose the following conundrum: When you pick up a cup of water, is it your *hand* that picks it up?

Most people, of course, go with the naive popular answer: "Yes."

Recently, however, scientists have made a stunning discovery: It's not your *hand* that holds the cup, it's actually your fingers, thumb, and palm.

Yes, I know! I was shocked too. But it seems that after scientists measured the forces exerted on the cup by each of your fingers, your thumb, and your palm, they found there was no force left over—so the force exerted by your *hand* must be zero.

The theme here is that, if you can *see how* (not just *know that*) a higher level reduces to a lower one, they will not seem like separate things within your map; you will be able to *see* how silly it is to think that your fingers could be in one place, and your hand somewhere else; you will be able to *see* how silly it is to argue about whether it is your hand picks up the cup, or your fingers.

---

1. [http://lesswrong.com/lw/p2/hand\\_vs\\_fingers/](http://lesswrong.com/lw/p2/hand_vs_fingers/)

2. Page 699, 'Reductionism'.

3. Page 703, 'Explaining vs. Explaining Away'.

4. Page 707, 'Fake Reductionism'.

5. Page 699, 'Reductionism'.

6. Page 685, 'Mind Projection Fallacy'.

7. Page 715, 'Joy in the Merely Real'.

8. Page 699, 'Reductionism'.

The operative word is “see”, as in concrete visualization. Imagining your hand causes you to imagine the fingers and thumb and palm; conversely, imagining fingers and thumb and palm causes you to identify a hand in the mental picture. Thus the high level of *your map* and the low level of *your map* will be tightly bound together *in your mind*.

In reality, of course, the levels are bound together even tighter than that—bound together by the tightest possible binding: physical identity. You can *see* this: You can *see* that saying (1) “hand” or (2) “fingers and thumb and palm”, does not refer to different *things*, but different *points of view*.

But suppose you lack the knowledge to so tightly bind together the levels of your map. For example, you could have a “hand scanner” that showed a “hand” as a dot on a map (like an old-fashioned radar display), and similar scanners for fingers/thumbs/palms; then you would see a cluster of dots around the hand, but you would be able to *imagine* the hand-dot moving off from the others. So, even though the physical reality of the hand (that is, the thing the dot corresponds to) was identical with / strictly composed of the physical realities of the fingers and thumb and palm, you would not be able to see this fact; even if someone told you, or you guessed from the correspondence of the dots, you would only *know* the fact of reduction, not *see* it. You would still be able to *imagine* the hand dot moving around independently, even though, if the physical makeup of the sensors were held constant, it would be physically impossible for this to actually happen.

Or, at a still lower level of binding, people might just tell you “There’s a hand over there, and some fingers over there”—in which case you would know little more than a Good-Old-Fashioned AI representing the situation using suggestively named LISP tokens. There wouldn’t be anything *obviously* contradictory about asserting:

```
|—Inside (Room, Hand)
|~Inside (Room, Fingers)
```

because you would not possess the *knowledge*

$$|\neg \text{Inside}(x, \text{Hand}) \rightarrow \text{Inside}(x, \text{Fingers})$$

None of this says that a hand can actually detach its existence from your fingers and crawl, ghostlike, across the room; it just says that a Good-Old-Fashioned AI with a propositional representation may not *know* any better. The map is not the territory.

In particular, you shouldn't draw too many conclusions from how it seems *conceptually possible*, in the mind of some specific conceiver, to separate the hand from its constituent elements of fingers, thumb, and palm. Conceptual possibility is not the same as logical possibility or physical possibility.

It is *conceptually possible to you* that 235757 is prime, because you don't know any better. But it isn't *logically possible* that 235757 is prime; if you were logically omniscient, 235757 would be obviously composite (and you would know the factors). That that's why we have the notion of impossible possible worlds, so that we can put probability distributions on propositions that may or may not be *in fact*<sup>9</sup> logically impossible.

And you can imagine philosophers who criticize “eliminative fingerists” who contradict the direct facts of experience—we can *feel* our hand holding the cup, after all—by suggesting that “hands” *don't really exist*, in which case, obviously, the cup would fall down. And philosophers who suggest “appendigital bridging laws” to explain how a particular configuration of fingers, evokes a hand into existence—with the note, of course, that while our world contains those particular appendigital bridging laws, the laws could have been conceivably different, and so are not in any sense *necessary facts*, etc.

All of these are cases of Mind Projection Fallacy, and what I call “naive philosophical realism”—the confusion of philosophical intuitions for direct, veridical information about reality. Your inability to imagine something is just a computational fact about what your brain can or can't imagine. Another brain might work differently.

---

9. Page 695, ‘Qualitatively Confused’.

## 25. Angry Atoms<sup>1</sup>

### Followup to: Hand vs. Fingers<sup>2</sup>

Fundamental physics—quarks ‘n stuff—is far removed from the levels we can *see*,<sup>3</sup> like hands and fingers. At best, you can know how to replicate the experiments which show that your hand (like everything else) is composed of quarks, and you may know how to derive a few equations for things like atoms and electron clouds and molecules.

At worst, the existence of quarks beneath your hand may just be something you were told<sup>4</sup>. In which case it’s questionable in one what sense you can be said to “know” it at all, even if you repeat back the same word “quark” that a physicist would use to convey knowledge to another physicist.

Either way, you can’t actually *see* the identity between levels—no one has a brain large enough to *visualize* avogadros of quarks and recognize a hand-pattern in them.

But we at least understand what hands *do*. Hands push on things, exert forces on them. When we’re told about atoms, we visualize little billiard balls bumping into each other. This makes it seem obvious that “atoms” can push on things too, by bumping into them.

Now this notion of atoms is not quite correct. But so far as *human imagination* goes, it’s relatively easy to imagine our hand being made up of a little galaxy of swirling billiard balls, pushing on things when our “fingers” touch them. Democritus imagined this 2400 years ago, and there was a time, roughly 1803-1922<sup>5</sup>, when Science thought he was right.

But what about, say, anger?

How could little billiard balls be angry? Tiny frowny faces on the billiard balls?

Put yourself in the shoes of, say, a hunter-gatherer—someone who may not even have a notion of writing, let alone the notion of

---

1. [http://lesswrong.com/lw/p3/angry\\_atoms/](http://lesswrong.com/lw/p3/angry_atoms/)

2. Page 751, ‘Hand vs. Fingers’.

3. Page 751, ‘Hand vs. Fingers’.

4. Page 140, ‘Guessing the Teacher’s Password’.

5. <http://en.wikipedia.org/wiki/Atom>

using base matter to perform computations—someone who has no idea that such a thing as neurons exist. Then you can imagine the *functional* gap that your ancestors might have perceived between billiard balls and “Grrr! Aaarg!”

Forget about subjective experience for the moment, and consider the sheer *behavioral* gap between anger and billiard balls. The difference between what little billiard balls *do*, and what anger makes people *do*. Anger can make people raise their fists and hit someone—or say snide things behind their backs—or plant scorpions in their tents at night. Billiard balls just push on things.

Try to put yourself in the shoes of the hunter-gatherer who’s never had the “Aha!” of information-processing. Try to avoid hindsight bias<sup>6</sup> about things like neurons and computers. Only then will you be able to see the uncrossable explanatory gap:

How can you explain angry behavior in terms of billiard balls?

Well, the *obvious* materialist conjecture is that the little billiard balls push on your arm and make you hit someone, or push on your tongue so that insults come out.

But how do the little billiard balls know how to do this—or how to guide your tongue and fingers through long-term plots—if they aren’t angry themselves?

And besides, if you’re not seduced by—gasp!—scientism, you can see from a first-person perspective that this explanation is obviously false. Atoms can push on your arm, but they can’t make you *want* anything.

Someone may point out that drinking wine can make you angry. But who says that wine is made exclusively of little billiard balls? Maybe wine just contains a potency of angerness.

Clearly, reductionism is just a flawed notion.

(The novice goes astray and says “The art failed me”; the master goes astray and says “I failed my art.”)

What does it take to cross this gap? It’s not just the idea of “neurons” that “process information”—if you say only this and nothing more, it just inserts a magical, unexplained level-crossing rule into your model, where you go from billiards to thoughts.

---

6. Page 170, ‘Failing to Learn from History’.

But an Artificial Intelligence programmer who knows how to create a chess-playing program out of base matter, has taken a *genuine* step toward crossing the gap. If you understand concepts like consequentialism<sup>7</sup>, backward chaining, utility functions, and search trees<sup>8</sup>, you can make merely causal/mechanical systems compute plans.

The trick goes something like this: For each possible chess move, compute the moves your opponent could make, then your responses to those moves, and so on; evaluate the furthest position you can see using some local algorithm (you might simply count up the material); then trace back using minimax<sup>9</sup> to find the best move on the current board; then make that move.

More generally: If you have chains of causality inside the mind that have a kind of mapping—a mirror, an echo—to what goes on in the environment, then you can run a utility function over the end products of imagination, and find an action that achieves something which the utility function rates highly, and output that action. It is not necessary for the chains of causality inside the mind, that are similar to the environment, to be made out of billiard balls that have little auras of intentionality. Deep Blue’s transistors do not need little chess pieces carved on them, in order to work. See also The Simple Truth<sup>10</sup>.

All this is still tremendously oversimplified, but it should, at least, reduce the apparent length of the gap. If you can understand all that, you can see how a planner built out of base matter can be influenced by alcohol to output more angry behaviors. The billiard balls in the alcohol push on the billiard balls making up the utility function.

But even if you know how to write small AIs, you can’t *visualize* the level-crossing between transistors and chess. There are too many transistors, and too many moves to check.

Likewise, even if you knew all the facts of neurology, you would not be able to *visualize* the level-crossing between neurons and anger—let alone the level-crossing between atoms and anger. Not

---

7. [http://lesswrong.com/lw/l4/terminal\\_values\\_and\\_instrumental\\_values/](http://lesswrong.com/lw/l4/terminal_values_and_instrumental_values/)

8. <http://en.wikipedia.org/wiki/Minimax>

9. <http://en.wikipedia.org/wiki/Minimax>

10. <http://yudkowsky.net/bayes/truth.html>

the way you can visualize a hand consisting of fingers, thumb, and palm.

And suppose a cognitive scientist just flatly tells<sup>11</sup> you “Anger is hormones”? Even if you repeat back the words, it doesn’t mean you’ve crossed the gap. You may believe you believe it<sup>12</sup>, but that’s not the same as understanding what little billiard balls have to do with wanting to hit someone.

So you come up with interpretations like, “Anger is *mere* hormones, it’s caused by little molecules, so it must not be justified in any moral sense—that’s why you should learn to control your anger.”

Or, “There isn’t really any such thing as anger—it’s an illusion, a quotation with no referent, like a mirage of water in the desert, or looking in the garage for a dragon and not finding one.”

These are both tough pills to swallow (not that you *should* swallow them) and so it is a good easier to profess<sup>13</sup> them than to believe them.

I think this is what non-reductionists/non-materialists think they are criticizing when they criticize reductive materialism.

But materialism isn’t that easy. It’s not as cheap as saying, “Anger is made out of atoms—there, now I’m done.” That wouldn’t explain how to get from billiard balls to hitting. You need the specific insights of computation, consequentialism, and search trees before you can start to close the explanatory gap.

All this was a relatively easy example *by modern standards*, because I restricted myself to talking about angry *behaviors*. Talking about outputs doesn’t require you to appreciate how an algorithm feels from inside<sup>14</sup> (cross a first-person/third-person gap) or dissolve a wrong question<sup>15</sup> (untangle places where the interior of your own mind runs skew to reality).

Going from material substances that bend and break, burn and fall, push and shove, to angry *behavior*, is just a practice problem by the standards of modern philosophy. But it is an *important* practice problem. It can only be fully appreciated, if you realize how *hard* it would have been to solve before writing was invented. There was

11. Page 707, ‘Fake Reductionism’.

12. Page 108, ‘Belief in Belief’.

13. Page 144, ‘Science as Attire’.

14. Page 223, ‘How An Algorithm Feels From Inside’.

15. Page 679, ‘Wrong Questions’.

once an explanatory gap here—though it may not seem that way in hindsight<sup>16</sup>, now that it's been bridged for generations.

Explanatory gaps can be crossed, if you accept help from science, and don't trust the view from the interior of your own mind.

---

16. Page 170, 'Failing to Learn from History'.

## 26. Heat vs. Motion<sup>1</sup>

### Followup to: Angry Atoms<sup>2</sup>

After yesterday's post, it occurred to me that there's a much simpler example of reductionism jumping a gap of apparent-difference-in-kind: the reduction of heat to motion.

Today, the equivalence of heat and motion may seem too obvious<sup>3</sup> in hindsight<sup>4</sup>—everyone says<sup>5</sup> that “heat is motion”, therefore, it can't be a “weird<sup>6</sup>” belief.

But there was a time when the kinetic theory of heat<sup>7</sup> was a highly controversial scientific hypothesis, contrasting to belief in a caloric fluid<sup>8</sup> that flowed from hot objects to cold objects. Still earlier, the main theory of heat was “Phlogiston!<sup>9</sup>”

Suppose you'd *separately* studied kinetic theory and caloric theory. You now know something about kinetics: collisions, elastic rebounds, momentum, kinetic energy, gravity, inertia, free trajectories. Separately, you know something about heat: Temperatures, pressures, combustion, heat flows, engines, melting, vaporization.

Not only is this state of knowledge a plausible one, it is the state of knowledge possessed by e.g. Sadi Carnot, who, working strictly from within the caloric theory of heat, developed the principle of the Carnot cycle—a heat engine of maximum efficiency, whose existence implies the second law of thermodynamics<sup>10</sup>. This in 1824, when kinetics was a highly developed science.

Suppose, like Carnot, you know a great deal about kinetics, and a great deal about heat, as *separate* entities. Separate entities of knowledge, that is: your brain has separate filing baskets for beliefs about kinetics and beliefs about heat. But from the inside<sup>11</sup>, this state of knowledge *feels* like living in a world of moving things and

---

1. [http://lesswrong.com/lw/p4/heat\\_vs\\_motion/](http://lesswrong.com/lw/p4/heat_vs_motion/)

2. Page 754, ‘Angry Atoms’.

3. Page 170, ‘Failing to Learn from History’.

4. Page 136, ‘Hindsight Devalues Science’.

5. Page 451, ‘Cached Thoughts’.

6. Page 432, ‘Cultish Countercultishness’.

7. [http://en.wikipedia.org/wiki/Theory\\_of\\_heat](http://en.wikipedia.org/wiki/Theory_of_heat)

8. <http://en.wikipedia.org/wiki/Caloric>

9. Page 147, ‘Fake Causality’.

10. [http://lesswrong.com/lw/o5/the\\_second\\_law\\_of\\_thermodynamics\\_and\\_engines\\_of/](http://lesswrong.com/lw/o5/the_second_law_of_thermodynamics_and_engines_of/)

11. Page 223, ‘How An Algorithm Feels From Inside’.

hot things, a world where motion and heat are independent properties of matter.

Now a Physicist From The Future comes along and tells you: “Where there is heat, there is motion, and vice versa. That’s why, for example, rubbing things together makes them hotter.”

There are (at least) two possible interpretations you could attach to this statement, “Where there is heat, there is motion, and vice versa.”

First, you could suppose that heat and motion exist separately—that the caloric theory is correct—but that among our universe’s physical laws is a “bridging law” which states that, where objects are moving quickly, caloric will come into existence. And conversely, another bridging law says that caloric can exert pressure on things and make them move, which is why a hotter gas exerts more pressure on its enclosure (thus a steam engine can use steam to drive a piston).

Second, you could suppose that heat and motion are, in some as-yet-mysterious sense, *the same thing*.

“Nonsense,” says Thinker 1, “the words ‘heat’ and ‘motion’ have two different meanings; that is why we have two different words. We know how to determine when we will call an observed phenomenon ‘heat’—heat can melt things, or make them burst into flame. We know how to determine when we will say that an object is ‘moving quickly’—it changes position; and when it crashes, it may deform, or shatter. Heat is concerned with change of substance; motion, with change of position and shape. To say that these two words have the same meaning is simply to confuse yourself.”

“Impossible,” says Thinker 2. “It may be that, in our world, heat and motion are associated by bridging laws, so that it is a law of physics that motion creates caloric, and vice versa. But I can easily imagine a world where rubbing things together does *not* make them hotter, and gases *don’t* exert more pressure at higher temperatures. Since there are possible worlds where heat and motion are not associated, they must be different properties—this is true *a priori*. ”

Thinker 1 is confusing the quotation and the referent<sup>12</sup>.  $2 + 2 = 4$ , but “ $2 + 2$ ” ≠ “4”. The string “ $2 + 2$ ” contains 5 characters (including whitespace) and the string “4” contains only 1 character. If you

---

<sup>12</sup>. Page 692, ‘The Quotation is not the Referent’.

type the two strings into a Python interpreter, they yield the same output,—> 4. So you can't conclude, from looking at the strings “ $2 + 2$ ” and “ $4$ ”, that just because the strings are different, they must have different “meanings” relative to the Python Interpreter.

The words “heat” and “kinetic energy” can be said to “refer to” the same thing, even before we *know* how heat reduces to motion, in the sense that we don't know yet what the reference is, but the references are in fact the same. You might imagine an Idealized Omniscient Science Interpreter that would give the same output when we typed in “heat” and “kinetic energy” on the command line.

I talk about the Science Interpreter to emphasize that, to dereference the pointer, you've got to step outside cognition. The end result of the dereference is something out there in reality, not in anyone's mind. So you can *say* “real referent” or “actual referent”, but you can't *evaluate* the words locally, from the inside of your own head. You can't reason using the actual heat-referent—if you thought using *real heat*, thinking “1 million Kelvin” would vaporize your brain. But, by forming a belief about your belief about heat, you can talk *about* your belief about heat, and say things like “It's possible that my belief about heat doesn't much resemble *real heat*.” You can't actually perform that comparison right there in your own mind, but you can talk *about* it.

Hence you can say, “My beliefs about heat and motion are not the same beliefs, but it's possible that actual heat and actual motion are the same thing.” It's just like being able to acknowledge that “the morning star” and “the evening star” might be the same planet, while also understanding that you can't determine this just by examining your beliefs—you've got to haul out the telescope.

Thinker 2's mistake follows similarly. A physicist told him, “Where there is heat, there is motion” and P2 mistook this for a statement of *physical law*: The presence of caloric *causes* the existence of motion. What the physicist really means is more akin to an *inferential rule*: Where you are told there is “heat”, deduce the presence of “motion”.

From this basic projection of a multilevel model into a multi-level reality follows another, distinct error: the conflation of conceptual possibility with logical possibility. To Sadi Carnot, it is *conceivable* that there could be another world where heat and mo-

tion are not associated. To Richard Feynman, armed with specific knowledge of how to derive equations about heat from equations about motion, this idea is not only inconceivable, but so wildly inconsistent as to make one's head explode.

I should note, in fairness to philosophers, that there are philosophers who have said these things. For example, Hilary Putnam, writing<sup>13</sup> on the “Twin Earth” thought experiment:

Once we have discovered that water (in the actual world) is H<sub>2</sub>O, *nothing counts as a possible world in which water isn't H<sub>2</sub>O*. In particular, if a “logically possible” statement is one that holds in some “logically possible world”, *it isn't logically possible that water isn't H<sub>2</sub>O*.

On the other hand, we can perfectly well imagine having experiences that would convince us (and that would make it rational to believe that) water *isn't* H<sub>2</sub>O. In that sense, it is conceivable that water isn't H<sub>2</sub>O. It is conceivable but it isn't logically possible! Conceivability is no proof of logical possibility.

It appears to me that “water” is being used in two different senses in these two paragraphs—one in which the word “water” refers to what we type into the Science Interpreter, and one in which “water” refers to what we get out of the Science Interpreter when we type “water” into it. In the first paragraph, Hilary seems to be saying that after we do some experiments and find out that water is H<sub>2</sub>O, water becomes automatically redefined to mean H<sub>2</sub>O. But you could coherently hold a different position about whether the word “water” now means “H<sub>2</sub>O” or “whatever is *really* in that bottle next to me”, so long as you use your terms consistently.

I believe the above has already been said as well? Anyway...

It is quite possible for there to be only *one* thing out-there-in-the-world, but for it to take on sufficiently different forms, and for you yourself to be sufficiently ignorant of the reduction, that it feels like living in a world containing two entirely different things.

---

<sup>13</sup> <http://books.google.com/books?id=cEPNtxAoYPAC&pg=PA18&lpg=PA18&dq=logically+possible&source=web&ots=R9zyWv03Uz&sig=FjyRRkSR88&hl=en>

Knowledge concerning these two different phenomena may taught in two different classes, and studied by two different academic fields, located in two different buildings of your university.

You've got to put yourself quite a ways back, into a historically realistic frame of mind, to remember how *different* heat and motion once seemed. Though, depending on how much you know today, it may not be as hard as all that, if you can look past the pressure of conventionality (that is, "heat is motion" is an un-weird belief, "heat is not motion" is a weird belief). I mean, suppose that tomorrow the physicists stepped forward and said, "Our popularizations of science have always contained one lie<sup>14</sup>. Actually, heat has nothing to do with motion." Could you *prove* they were wrong?

Saying "Maybe heat and motion are the same thing!" is easy. The difficult part is explaining *how*. It takes a great deal of detailed knowledge to get yourself to the point where you can no longer *conceive* of a world in which the two phenomena go separate ways. Reduction isn't cheap, and that's why it buys so much.

Or maybe you could say: "Reductionism is easy, reduction is hard." But it does kinda help to be a reductionist, I think, when it comes time to go looking for a reduction.

---

14. <http://www.overcomingbias.com/2008/02/my-favorite-lia.html>

## 27. Brain Breakthrough! It's Made of Neurons!<sup>1</sup>

In an amazing breakthrough<sup>2</sup>, a multinational team of scientists led by Nobel laureate Santiago Ramón y Cajal announced that the brain is composed of a *ridiculously* complicated network of tiny cells connected to each other by infinitesimal threads and branches.

The multinational team—which also includes the famous technician Antonie van Leeuwenhoek, and possibly Imhotep, promoted to the Egyptian god of medicine—issued this statement:

“The present discovery culminates years of research indicating that the convoluted squishy thing inside our skulls is even more complicated than it looks. Thanks to Cajal’s application of a new staining technique invented by Camillo Golgi, we have learned that this structure is not a continuous network like the blood vessels of the body, but is actually composed of many tiny cells, or “neurons”, connected to one another by even more tiny filaments.

“Other extensive evidence, beginning from Greek medical researcher Alcmaeon and continuing through Paul Broca’s research on speech deficits, indicates that the brain is the seat of reason.

“Nemesius, the Bishop of Emesia, has previously argued that brain tissue is too earthy to act as an intermediary between the body and soul, and so the mental faculties are located in the ventricles of the brain. However, if this is correct, there is no reason why this organ should turn out to have an immensely complicated internal composition.

“Charles Babbage has independently suggested that many small mechanical devices could be collected into an ‘Analytical Engine’, capable of performing activities, such as arithmetic, which are widely believed to require thought. The work of Luigi Galvani and Hermann von Helmholtz suggests that the activities of neurons are electrochemical in nature, rather than mechanical pressures as previously believed. Nonetheless, we think an analogy with Babbage’s ‘Analytical Engine’ suggests that a vastly complicated network of neurons could similarly exhibit thoughtful properties.

---

1. [http://lesswrong.com/lw/p5/brain\\_breakthrough\\_its\\_made\\_of\\_neurons/](http://lesswrong.com/lw/p5/brain_breakthrough_its_made_of_neurons/)

2. Page 734, ‘Amazing Breakthrough Day: April 1st’.

"We have found an enormously complicated material system located where the mind should be. The implications are shocking, and must be squarely faced. We believe that the present research offers strong experimental evidence that Benedictus Spinoza was correct, and René Descartes wrong: Mind and body are of one substance.

"In combination with the work of Charles Darwin showing how such a complicated organ could, in principle, have arisen as the result of processes not themselves intelligent, the bulk of scientific evidence now seems to indicate that intelligence is ontologically non-fundamental and has an extended origin in time. This strongly weighs against theories which assign mental entities an ontologically fundamental or causally primal status, including all religions ever invented.

"Much work remains to be done on discovering the specific identities between electrochemical interactions between neurons, and thoughts. Nonetheless, we believe our discovery offers the promise, though not yet the realization, of a full scientific account of thought. The problem may now be declared, if not solved, then solvable."

We regret that Cajal and most of the other researchers involved on the Project are no longer available for comment.

## 28. Reductive Reference<sup>1</sup>

**Followup to:** Reductionism<sup>2</sup>, Explaining vs. Explaining Away<sup>3</sup>, Hand vs. Fingers<sup>4</sup>, Heat vs. Motion<sup>5</sup>

The reductionist thesis (as I formulate it) is that human minds, for reasons of efficiency, use a multi-level map in which we separately *think* about things like “atoms” and “quarks”, “hands” and “fingers”, or “heat” and “kinetic energy”. Reality itself, on the other hand, is single-level in the sense that it does not seem to contain atoms as *separate, additional, causally efficacious entities over and above* quarks.

Sadi Carnot formulated the (precursor to) the second law of thermodynamics using the caloric theory of heat, in which heat was just a fluid that flowed from hot things to cold things, produced by fire, making gases expand—the effects of heat were studied separately from the science of kinetics, considerably before the reduction took place. If you’re trying to design a steam engine, the effects of all those tiny vibrations and collisions which we name “heat” can be summarized into a much simpler description than the full quantum mechanics of the quarks. Humans compute efficiently, thinking of only significant effects on goal-relevant quantities.

But reality itself does seem to use the full quantum mechanics of the quarks. I once met a fellow<sup>6</sup> who thought that if you used General Relativity to compute a low-velocity problem, like an artillery shell, GR would give you the *wrong answer*—not just a slow answer, but an *experimentally wrong* answer—because at low velocities, artillery shells are governed by Newtonian mechanics, not GR. This is exactly how physics does *not* work. Reality just seems to go on crunching through General Relativity, even when it only makes a difference at the fourteenth decimal place, which a human would regard as a huge waste of computing power. Physics does it with brute force. No one has *ever* caught physics simplifying its calcu-

---

1. [http://lesswrong.com/lw/p6/reductive\\_reference/](http://lesswrong.com/lw/p6/reductive_reference/)

2. Page 699, ‘Reductionism’.

3. Page 703, ‘Explaining vs. Explaining Away’.

4. Page 751, ‘Hand vs. Fingers’.

5. Page 759, ‘Heat vs. Motion’.

6. Page 699, ‘Reductionism’.

lations—or if someone did catch it, the Matrix Lords erased the memory afterward.

Our map, then, is very much unlike the territory; our maps are multi-level, the territory is single-level. Since the representation is so incredibly unlike the referent, in what sense can a belief like “I am wearing socks” be called *true*, when in reality itself, there are only quarks?

In case you’ve forgotten what the word “true” means, the classic definition was given by Alfred Tarski:

The statement “snow is white” is *true* if and only if snow is white.

In case you’ve forgotten what the difference is between the statement “I believe ‘snow is white’” and “‘Snow is white’ is true”, see here<sup>7</sup>. Truth can’t be evaluated *just* by looking inside your own head—if you want to know, for example, whether “the morning star = the evening star”, you need a telescope; it’s not enough just to look at the beliefs themselves.

This is the point missed by the postmodernist folks screaming, “But how do you *know* your beliefs are true?” When you do an experiment, you actually *are* going outside your own head. You’re engaging in a complex interaction whose outcome is causally determined by the thing you’re reasoning about, not just your beliefs about it. I once defined “reality” as follows<sup>8</sup>:

Even when I have a simple hypothesis, strongly supported by all the evidence I know, sometimes I’m still surprised. So I need different names for the thingies that determine my predictions and the thingy that determines my experimental results. I call the former thingies ‘belief’, and the latter thingy ‘reality’.”

The interpretation of your experiment still depends on your prior beliefs. I’m not going to talk, for the moment, about Where Priors Come From, because that is not the subject of this blog post. My point is that truth refers to an *ideal* comparison between a belief

---

7. Page 695, ‘Qualitatively Confused’.

8. <http://yudkowsky.net/bayes/truth.html>

and reality. Because we understand that planets are distinct from beliefs about planets, we can design an experiment to test whether the belief “the morning star and the evening star are the same planet” is *true*. This experiment will involve telescopes, not just introspection, because we understand that “truth” involves comparing an internal belief to an external fact; so we use an instrument, the telescope, whose perceived behavior we believe to depend on the external fact of the planet.

Believing that the telescope helps us evaluate the “truth” of “morning star = evening star”, relies on our prior beliefs about the telescope interacting with the planet. Again, I’m not going to address that in this particular blog post, except to quote one of my favorite Raymond Smullyan lines: “If the more sophisticated reader objects to this statement on the grounds of its being a mere tautology, then please at least give the statement credit for not being inconsistent.” Similarly, I don’t see the use of a telescope as circular logic, but as reflective coherence; for every systematic way of arriving at truth, there ought to be a rational explanation for how it works.

The question on the table is what it *means* for “snow is white” to be *true*, when, in reality, there are just quarks.

There’s a certain pattern of neural connections making up your beliefs about “snow” and “whiteness”—we believe this, but we do not know, and cannot concretely visualize, the actual neural connections. Which are, themselves, embodied in a pattern of quarks even less known. Out there in the world, there are water molecules whose temperature is low enough that they have arranged themselves in tiled repeating patterns; they look nothing like the tangles of neurons. In what sense, comparing one (ever-fluctuating) pattern of quarks to the other, is the belief “snow is white” *true*?

Obviously, neither I nor anyone else can offer an Ideal Quark Comparer Function that accepts a quark-level description of a neurally embodied belief (including the surrounding brain) and a quark-level description of a snowflake (and the surrounding laws of optics), and outputs “true” or “false” over “snow is white”. And who says the fundamental level is *really* about particle fields?

On the other hand, throwing out all beliefs because they aren’t written as gigantic unmanageable specifications about quarks we

can't even see... doesn't seem like a very prudent idea. Not the best way to optimize our goals.<sup>9</sup>

It seems to me that a word like “snow” or “white” can be taken as a kind of promissory note—not a *known* specification of exactly which physical quark configurations count as “snow”, but, nonetheless, there are things you call snow and things you don’t call snow<sup>10</sup>, and even if you got a few items wrong (like plastic snow), an Ideal Omniscent Science Interpreter would see a tight cluster in the center<sup>11</sup> and redraw the boundary<sup>12</sup> to have a simpler definition<sup>13</sup>.

In a single-layer universe whose bottom layer is unknown, or uncertain, or just too large to talk about, the concepts in a multi-layer mind can be said to represent a kind of promissory note—we don’t know *what* they correspond to, out there. But it seems to us that we can distinguish positive from negative cases, in a predictively productive way<sup>14</sup>, so we think—perhaps in a fully general sense—that there is *some* difference of quarks, *some* difference of configurations at the fundamental level, which explains the differences that feed into our senses, and ultimately result in our saying “snow” or “not snow”.

I see this white stuff, and it is the same on several occasions, so I hypothesize a stable latent cause in the environment—I give it the name “snow”; “snow” is then a promissory note referring to a believed-in simple boundary that could be drawn around the unseen causes of my experience.

Hilary Putnam’s “Twin Earth” thought experiment, where water is not H<sub>2</sub>O but some strange other substance denoted XYZ, otherwise behaving much like water, and the subsequent philosophical debate, helps to highlight this issue. “Snow” doesn’t have a logical definition known to us—it’s more like an empirically determined pointer to a logical definition. This is true even if you believe that snow is ice crystals is low-temperature tiled water molecules. The water molecules are made of quarks. What if quarks turn out to be made of something else? What *is* a snowflake, then? You don’t know—but it’s still a snowflake, not a fire hydrant.

---

9. Page 644, ‘Something to Protect’.

10. Page 200, ‘Extensions and Intensions’.

11. Page 209, ‘The Cluster Structure of Thingspace’.

12. Page 267, ‘Where to Draw the Boundary?’.

13. Page 282, ‘Superexponential Conceptspace, and Simple Words’.

14. Page 275, ‘Mutual Information, and Density in Thingspace’.

And of course, these very paragraphs I have just written, are likewise far above the level of quarks. “Sensing white stuff, visually categorizing it, and thinking ‘snow’ or ‘not snow’”—this is also talking very far above the quarks. So my meta-beliefs are also promissory notes, for things that an Ideal Omniscient Science Interpreter might know about which configurations of the quarks (or whatever) making up my brain, correspond to “believing ‘snow is white’”.

But then, the entire grasp that we have upon reality, is made up of promissory notes of this kind. So, rather than calling it circular, I prefer to call it self-consistent.

This can be a bit unnerving—maintaining a precarious epistemic perch, in both object-level beliefs and reflection, far above a huge unknown underlying fundamental reality, and hoping one doesn’t fall off.

On reflection, though, it’s hard to see how things could be any other way.

So at the end of the day, the statement “reality does not contain hands as fundamental, additional, separate causal entities, over and above quarks” is not the same statement as “hands do not exist” or “I don’t have any hands”. There are no *fundamental* hands; hands are made of fingers, palm, and thumb, which in turn are made of muscle and bone, all the way down to elementary particle fields, which are the fundamental causal entities, so far as we currently know.

This is not the same as saying, “there are no ‘hands’.” It is not the same as saying, “the word ‘hands’ is a promissory note that will never be paid, because there is no empirical cluster that corresponds to it”; or “the ‘hands’ note will never be paid, because it is logically impossible to reconcile its supposed characteristics”; or “the statement ‘humans have hands’ refers to a sensible state of affairs, but reality is not in that state”.

Just: There are patterns that exist *in* reality where we see “hands”, and these patterns have something in common, but they are not fundamental.

If I *really* had no hands—if reality suddenly transitioned to be in a state that we would describe as “Eliezer has no hands”—reality would shortly thereafter correspond to a state we would describe as “Eliezer screams as blood jets out of his wrist stumps”.

And this is *true*, even though the above paragraph hasn't specified any quark positions.

The previous sentence is likewise meta-true.

The map is multilevel, the territory is single-level. This doesn't mean that the higher levels "don't exist", like looking in your garage for a dragon and finding nothing there, or like seeing a mirage in the desert and forming an expectation of drinkable water when there is nothing to drink. The higher levels of your map are not *false*, without referent; they have referents *in* the single level of physics. It's not that the wings of an airplane unexist—then the airplane would drop out of the sky. The "wings of an airplane" exist *explicitly* in an engineer's multilevel model of an airplane, and the wings of an airplane exist *implicitly* in the quantum physics of the real airplane. Implicit existence is not the same as nonexistence. The exact description of this implicitness is not known to us—is not explicitly represented in our map. But this does not prevent our map from working, or even prevent it from being *true*.

Though it is a bit unnerving to contemplate that every single concept and belief in your brain, including these meta-concepts about how your brain works and why you can form accurate beliefs, are perched orders and orders of magnitude above reality...

## 29. Zombies! Zombies?<sup>1</sup>

<sup>2</sup>Your “zombie”, in the philosophical usage of the term, is putatively a being that is exactly like you in *every* respect—identical behavior, identical speech, identical brain; every atom and quark in *exactly* the same position, moving according to the same causal laws of motion—*except* that your zombie is not conscious.

It is furthermore claimed that if zombies are “possible” (a term over which battles are still being fought), then, purely from our knowledge of this “possibility”, we can deduce a priori that consciousness is extra-physical, in a sense to be described below; the standard term for this position is “epiphenomenalism”.

(For those unfamiliar with zombies, I emphasize that *this is not a strawman*. See, for example, the SEP entry on Zombies<sup>3</sup>. The “possibility” of zombies is accepted by a substantial fraction, possibly a majority, of academic philosophers of consciousness.)

I once read somewhere, “You are not the one who speaks your thoughts—you are the one who *bears* your thoughts”. In Hebrew, the word for the highest soul, that which God breathed into Adam, is N’Shama—“the hearer”.

If you conceive of “consciousness” as a purely passive listening, then the notion of a zombie initially seems easy to imagine. It’s someone who lacks the N’Shama, the hearer.

(Warning: Long post ahead. *Very* long 6,600-word post involving David Chalmers ahead. This may be taken as my demonstrative counterexample to Richard Chappell’s Arguing with Eliezer Part II<sup>4</sup>, in which Richard accuses me of not engaging with the complex arguments of real philosophers.)

---

1. [http://lesswrong.com/lw/p7/zombies\\_zombies/](http://lesswrong.com/lw/p7/zombies_zombies/)

2. <http://flickr.com/photos/doviende/38008649/>

3. <http://plato.stanford.edu/entries/zombies/>

4. <http://www.philosophyetc.net/2008/03/arguing-with-eliezer-part-ii.html>

When you open a refrigerator and find that the orange juice is gone, you think “Darn, I’m out of orange juice.” The sound of these words is probably represented in your auditory cortex, as though you’d heard someone else say it. (Why do I think this? Because native Chinese speakers can remember longer digit sequences than English-speakers. Chinese digits are all single syllables, and so Chinese speakers can remember around ten digits, versus the famous “seven plus or minus two” for English speakers. There appears to be a loop of repeating sounds back to yourself, a size limit on working memory in the auditory cortex, which is genuinely phoneme-based.)

Let’s suppose the above is correct; as a postulate, it should certainly present no problem for advocates of zombies. Even if humans are not like this, it seems easy enough to imagine an AI constructed this way (and imaginability is what the zombie argument is all about). It’s not only conceivable in principle, but quite possible in the next couple of decades, that surgeons will lay a network of neural taps over someone’s auditory cortex and read out their internal narrative. (Researchers have already tapped the lateral geniculate nucleus of a cat and reconstructed recognizable visual inputs.)

So your zombie, being physically identical to you down to the last atom, will open the refrigerator and form auditory cortical patterns for the phonemes “Darn, I’m out of orange juice”. On this point, epiphenomalists would willingly agree.

But, says the epiphenomenalist, in the zombie there is no one inside to *bear*; the inner listener is missing. The internal narrative is spoken, but unheard. You are not the one who speaks your thoughts, you are the one who hears them.

It seems a lot more straightforward (they would say) to make an AI that prints out some kind of internal narrative, than to show that an inner listener hears it.

The Zombie Argument is that if the Zombie World is *possible*—not necessarily physically possible in our universe, just “possible in theory”, or “imaginable”, or something along those lines—then consciousness must be extra-physical, something over and above mere atoms. Why? Because even if you somehow knew the positions of all the atoms in the universe, you would still have

be told, as a separate and additional fact, that people were conscious—that they had inner listeners—that we were not in the Zombie World, as seems *possible*.

Zombie-ism is not the same as dualism. Descartes thought there was a body-substance and a wholly different kind of mind-substance, but Descartes also thought that the mind-substance was a *causally active* principle, interacting with the body-substance, controlling our speech and behavior. Subtracting out the mind-substance from the human would leave a *traditional* zombie, of the lurching and groaning sort.

And though the Hebrew word for the innermost soul is N'Shama, that-which-hears, I can't recall hearing a rabbi arguing for the possibility of zombies. Most rabbis would probably be aghast at the idea that the divine part which God breathed into Adam *doesn't actually do anything*.

The technical term for the belief that consciousness is there, but has no effect on the physical world, is *epiphenomenalism*.

Though there are other elements to the zombie argument (I'll deal with them below), I think that the intuition of the passive listener is what first seduces people to zombie-ism. In particular, it's what seduces a lay audience to zombie-ism. The core notion is simple and easy to access: The lights are on but no one's home.

Philosophers are appealing to the intuition of the passive listener when they say "Of course the zombie world is imaginable; you know exactly what it would be like."

One of the great battles in the Zombie Wars is over what, exactly, is meant by saying that zombies are "possible". Early zombie-ist philosophers (the 1970s) just thought it was obvious that zombies were "possible", and didn't bother to define what sort of possibility was meant.

Because of my reading in mathematical logic, what instantly comes into my mind is logical possibility. If you have a collection of statements like  $(A \rightarrow B), (B \rightarrow C), (C \rightarrow \neg A)$  then the compound belief is *logically possible* if it has a *model*—which, in the simple case above, reduces to finding a value assignment to A, B, C that makes all of the statements  $(A \rightarrow B), (B \rightarrow C)$ , and  $(C \rightarrow \neg A)$  true. In this case,  $A=B=C=0$  works, as does  $A=0, B=C=1$  or  $A=B=0, C=1$ .

Something will *seem* possible—will seem “conceptually possible” or “imaginable”—if you can consider the collection of statements without *seeing* a contradiction. But it is, in general, a very hard problem to see contradictions *or* to find a full specific model! If you limit yourself to simple Boolean propositions of the form ((A or B or C) and (B or -C or D) and (D or -A or -C) ...), conjunctions of disjunctions of three variables, then this is a very famous problem called 3-SAT, which is one of the first problems ever to be proven NP-complete.

So just because you don’t see a contradiction in the Zombie World at first glance, it doesn’t mean that no contradiction is there. It’s like not seeing a contradiction in the Riemann Hypothesis at first glance. From conceptual possibility (“I don’t see a problem”) to *logical possibility* in the full technical sense, is a very great leap. It’s easy to make it an NP-complete leap, and with first-order theories you can make it superexponential. And it’s *logical* possibility of the Zombie World, not conceptual possibility, that is needed to suppose that a logically omniscient mind could know the positions of all the atoms in the universe, and yet need to be told as an *additional* non-entailed fact that we have inner listeners.

Just because you don’t see a contradiction *yet*, is no guarantee that you won’t see a contradiction in another 30 seconds. “All odd numbers are prime. Proof: 3 is prime, 5 is prime, 7 is prime...”

So let us ponder the Zombie Argument *a little longer*: Can we think of a counterexample to the assertion “Consciousness has no third-party-detectable causal impact on the world”?

If you close your eyes and concentrate on your inward awareness, you will begin to form thoughts, in your internal narrative, that go along the lines of “I am aware” and “My awareness is separate from my thoughts” and “I am not the one who speaks my thoughts, but the one who hears them” and “My stream of consciousness is not my consciousness” and “It seems like there is a part of me which I can imagine being eliminated without changing my outward behavior.”

You can even say these sentences out loud, as you meditate. In principle, someone with a super-fMRI could probably read the phonemes out of your auditory cortex; but saying it out loud re-

moves all doubt about whether you have entered the realms of testability and physical consequences.

This certainly seems like the inner listener is being *caught in the act of listening* by whatever part of you writes the internal narrative and flaps your tongue.

Imagine that a mysterious race of aliens visit you, and leave you a mysterious black box as a gift. You try poking and prodding the black box, but (as far as you can tell) you never succeed in eliciting a reaction. You can't make the black box produce gold coins or answer questions. So you conclude that the black box is causally inactive: "For all X, the black box doesn't do X." The black box is an effect, but not a cause; epiphenomenal; without causal potency. In your mind, you test this general hypothesis to see if it is true in some trial cases, and it seems to be true—"Does the black box turn lead to gold? No. Does the black box boil water? No."

But you can *see* the black box; it absorbs light, and weighs heavy in your hand. This, too, is part of the dance of causality. If the black box were *wholly* outside the causal universe, you couldn't see it; you would have no way to know it existed; you could not say, "Thanks for the black box." You didn't *think* of this counterexample, when you formulated the general rule: "All X: Black box doesn't do X". But it was there all along.

(Actually, the aliens left you *another* black box, this one *purely* epiphenomenal, and you haven't the slightest clue that it's there in your living room. That was their joke.)

If you can close your eyes, and sense yourself sensing—if you can be aware of yourself being aware, and think "I am aware that I am aware"—and say out loud, "I am aware that I am aware"—then your consciousness is not without effect on your internal narrative, or your moving lips. You can see yourself seeing, and your internal narrative reflects this, and so do your lips if you choose to say it out loud.

I have not seen the above argument written out that particular way—"the listener caught in the act of listening"—though it may well have been said before.

But it is a standard point<sup>5</sup>—which zombie-ist philosophers accept!—that the Zombie World's philosophers, being atom-by-atom

---

5. <http://plato.stanford.edu/entries/zombies/>

identical to our own philosophers, write identical papers about the philosophy of consciousness.

At this point, the Zombie World stops being an intuitive consequence of the idea of a passive listener.

Philosophers writing papers about consciousness would *seem* to be at least one effect of consciousness upon the world. You can argue clever reasons why this is not so, but you have to be clever.

You would intuitively suppose that if your inward awareness went away, this would change the world, in that your internal narrative would no longer say things like “There is a mysterious listener within me,” because the mysterious listener would be gone. It is usually right *after* you focus your awareness on your awareness, that your internal narrative says “I am aware of my awareness”, which suggests that if the first event never happened again, neither would the second. You can argue clever reasons why this is not so, but you have to be clever.

You can form a propositional belief that “Consciousness is without effect”, and not *see* any contradiction at first, if you don’t realize that talking about consciousness is an effect of being conscious. But once you see<sup>6</sup> the connection from the general rule that consciousness has no effect, to the specific implication that consciousness has no effect on how philosophers write papers about consciousness, zombie-ism stops being intuitive and starts requiring you to postulate strange things.

One strange thing you might postulate is that there’s a Zombie Master, a god within the Zombie World who surreptitiously takes control of zombie philosophers and makes them talk and write about consciousness.

A Zombie Master doesn’t seem impossible. Human beings often don’t sound all that coherent when talking about consciousness. It might not be that hard to fake their discourse, to the standards of, say, a human amateur talking in a bar. Maybe you could take, as a corpus, one thousand human amateurs trying to discuss consciousness; feed them into a non-conscious but sophisticated AI, better than today’s models but not self-modifying; and get back discourse about “consciousness” that sounded as sensible as most humans, which is to say, not very.

---

6. Page 751, ‘Hand vs. Fingers’.

But this speech about “consciousness” would not be spontaneous. It would not be produced *within* the AI. It would be a recorded imitation<sup>7</sup> of someone else talking. That is just a holodeck, with a central AI writing the speech of the non-player characters<sup>8</sup>. This is *not* what the Zombie World is about.

By supposition, the Zombie World is atom-by-atom identical to our own, except that the inhabitants lack consciousness. Furthermore, the atoms in the Zombie World move under the same laws of physics as in our own world. If there are “bridging laws” that govern *which configurations of atoms evoke consciousness*, those bridging laws are absent. But, by hypothesis, the difference is not experimentally detectable. When it comes to saying whether a quark zigs or zags or exerts a force on nearby quarks—anything experimentally measurable—the same physical laws govern.

The Zombie World has no *room* for a Zombie Master, because a Zombie Master has to control the zombie’s lips, and that control is, in principle, experimentally detectable. The Zombie Master moves lips, therefore it has observable consequences. There would be a point where an electron zags, instead of zigging, because the Zombie Master says so. (Unless the Zombie Master is actually *in* the world, as a pattern of quarks—but then the Zombie World is not atom-by-atom identical to our own, unless you think *this* world also contains a Zombie Master.)

When a philosopher in our world types, “I think the Zombie World is possible”, his fingers strike keys in sequence: Z-O-M-B-I-E. There is a chain of causality that can be traced back from these keystrokes: muscles contracting, nerves firing, commands sent down through the spinal cord, from the motor cortex—and then into less understood areas of the brain, where the philosopher’s internal narrative first began talking about “consciousness”.

And the philosopher’s zombie twin strikes the same keys, *for the same reason*, causally speaking. There is no cause within the chain of explanation for why the philosopher writes the way he does, which is not also present in the zombie twin. The zombie twin also has an internal narrative about “consciousness”, that a super-fMRI could read out of the auditory cortex. And whatever other thoughts, or

---

7. [http://lesswrong.com/lw/ly/artificial\\_addition/](http://lesswrong.com/lw/ly/artificial_addition/)

8. <http://yudkowsky.net/fiction/NPC.html>

other causes of any kind, led to that internal narrative, they are exactly the same in our own universe and in the Zombie World.

So you can't say that the philosopher is writing about consciousness *because of* consciousness, while the zombie twin is writing about consciousness *because of* a Zombie Master or AI chatbot. When you trace back the chain of causality behind the keyboard, to the internal narrative echoed in the auditory cortex, to the cause of the narrative, you must find the *same* physical explanation in our world as in the zombie world.

As the most formidable advocate of zombie-ism, David Chalmers, writes<sup>9</sup>:

Think of my zombie twin in the universe next door. He talks about conscious experience all the time—in fact, he seems obsessed by it. He spends ridiculous amounts of time hunched over a computer, writing chapter after chapter on the mysteries of consciousness. He often comments on the pleasure he gets from certain sensory qualia, professing a particular love for deep greens and purples. He frequently gets into arguments with zombie materialists, arguing that their position cannot do justice to the realities of conscious experience.

And yet he has no conscious experience at all! In his universe, the materialists are right and he is wrong. Most of his claims about conscious experience are utterly false. But there is certainly a physical or functional explanation of why he makes the claims he makes. After all, his universe is fully law-governed, and no events therein are miraculous, so there must be some explanation of his claims.

...Any explanation of my twin's behavior will equally count as an explanation of my behavior, as the processes inside his body are precisely mirrored by those inside

---

9. <http://books.google.com/books?id=oZZQHOJdAAC&dq=the+conscious+mind&pg=PP1&ots=qjv14gGUOR&sig=98dNHve8xZp-yltM37iN8mgqzA9&hl=en&prev=http://www.google.com/search?hl=en&q=the+conscious+mind&sa=X&oi=print&ct=title&cad=one-book-with-thumb-nail#PPA18o,M1>

mine. The explanation of his claims obviously does not depend on the existence of consciousness, as there is no consciousness in his world. It follows that the explanation of my claims is also independent of the existence of consciousness.

Chalmers is not arguing *against* zombies; those are his actual beliefs!

This paradoxical situation is at once delightful and disturbing. It is not obviously fatal to the nonreductive position, but it is at least something that we need to come to grips with...

I would seriously nominate this as the largest bullet ever bitten in the history of time. And that is a backhanded compliment to David Chalmers: A lesser mortal would simply fail to see the implications, or refuse to face them, or rationalize a reason it wasn't so.

Why would anyone bite a bullet that large? Why would anyone postulate unconscious zombies who write papers about consciousness for *exactly the same reason* that our own genuinely conscious philosophers do?

Not because of the first intuition I wrote about, the intuition of the passive listener. That intuition may say that zombies can drive cars or do math or even fall in love, but it doesn't say that zombies write philosophy papers about their passive listeners.

The zombie argument does not rest *solely* on the intuition of the passive listener. If this was all there was to the zombie argument, it would be dead by now, I think. The intuition that the "listener" can be eliminated without effect, would go away as soon as you realized that your internal narrative routinely *seems* to catch the listener in the act of listening.

No, the drive to bite *this* bullet comes from an entirely different intuition—the intuition that no matter how many atoms you add up, no matter how many masses and electrical charges interact with each other, they will never *necessarily* produce a subjective sensation of the mysterious **redness** of red. It may be a fact about our physical universe (Chalmers says) that putting such-and-such atoms into

such-and-such a position, *evokes* a sensation of **redness**; but if so, it is not a *necessary* fact, it is something to be explained above and beyond the motion of the atoms.

But if you consider the second intuition on its own, without the intuition of the passive listener, it is hard to see why it implies zombie-ism. Maybe there's just a *different kind of stuff*, apart from and additional to atoms, that is *not* causally passive—a soul that actually *does* stuff, a soul that plays a real causal role in why we write about “the mysterious redness of red”. Take out the soul, and... well, assuming you just don't fall over in a coma, you certainly won't write any more papers about consciousness!

This is the position taken by Descartes and most other ancient thinkers: The soul is of a different kind, but it *interacts* with the body. Descartes's position is technically known as *substance dualism*—there is a thought-stuff, a mind-stuff, and it is not like atoms; but it is causally potent, interactive, and leaves a visible mark on our universe.

Zombie-ists are *property dualists*—they don't believe in a *separate* soul; they believe that matter in our universe has *additional properties* beyond the physical.

“Beyond the physical”? What does that mean? It means the extra properties are there, but they don't influence the motion of the atoms, like the properties of electrical charge or mass. The extra properties are not experimentally detectable *by third parties*; you know you are conscious, from the *inside* of your extra properties, but no scientist can ever directly detect this from outside.

So the additional properties are there, but not causally active. The extra properties do not move atoms around, which is why they can't be detected by third parties.

And that's why we can (allegedly) imagine a universe just like this one, with all the atoms in the same places, but the extra properties missing, so that everything goes on the same as before, but no one is conscious.

The Zombie World may not be *physically* possible, say the zombie-ists—because it is a fact that all the matter in our universe has the extra properties, or obeys the bridging laws that evoke consciousness—but the Zombie World is *logically* possible: the bridging laws could have been different.

But, once you realize that conceivability is not the same as logical possibility, and that the Zombie World isn't even all that intuitive, why say that the Zombie World is logically possible?

Why, oh why, say that the extra properties are epiphenomenal and indetectable?

We can put this dilemma very sharply: Chalmers believes that there *is* something called consciousness, and this consciousness embodies the true and indescribable substance of the mysterious redness of red. It may be a property beyond mass and charge, but it's *there*, and it *is* consciousness. Now, having said the above, Chalmers furthermore specifies that this true stuff of consciousness is epiphenomenal, without causal potency—but *why say that?*

Why say that you could subtract this true stuff of consciousness, and leave all the atoms in the same place doing the same things? If that's true, we need some *separate* physical explanation for why Chalmers talks about “the mysterious redness of red”. That is, there exists both a mysterious redness of red, which is extra-physical, and *an entirely separate* reason, *within* physics, why Chalmers *talks* about the “mysterious redness of red”.

Chalmers does confess that these two things seem like they ought to be related, but really, why do you need both? Why not just pick one or the other?

Once you've postulated that there is a mysterious redness of red, why not just say that it interacts with your internal narrative and makes you talk about the “mysterious redness of red”?

Isn't Descartes taking the simpler approach, here? The *strictly* simpler approach?

Why postulate an extramaterial soul, *and then* postulate that the soul has no effect on the physical world, *and then* postulate a mysterious unknown material process that causes your internal narrative to talk about conscious experience?

Why not postulate the true stuff of consciousness which no amount of mere mechanical atoms can add up to, *and then*, having gone that far already, let this true stuff of consciousness have causal effects like making philosophers talk about consciousness?

I am not endorsing Descartes's view. But at least I can understand where Descartes is coming from. Consciousness seems

mysterious, so you postulate a mysterious stuff of consciousness<sup>10</sup>. Fine.

But now the zombie-ists postulate that this mysterious stuff *doesn't do anything*, so you need a *whole new* explanation for why you *say* you're conscious.

That isn't vitalism. That's something so bizarre that vitalists would spit out their coffee. "When fires burn, they release phlogiston<sup>11</sup>. But phlogiston doesn't have any experimentally detectable impact on our universe, so you'll have to go looking for a *separate* explanation of why a fire can melt snow." *What?*

Are property dualists under the impression that if they postulate a new *active* force, something that has a causal impact on observables, they will be sticking their necks out too far?

Me, I'd say that if you postulate a mysterious, separate, additional, inherently mental property of consciousness, above and beyond positions and velocities, then, at that point, you have *already* stuck your neck out as far as it can go. To postulate this stuff of consciousness, and then further postulate that it *doesn't do anything*—for the love of cute kittens, *why?*

There isn't even an obvious career motive. "Hi, I'm a philosopher of consciousness. My subject matter is the most important thing in the universe and I should get lots of funding? Well, it's nice of you to say so, but actually the phenomenon I study doesn't do anything whatsoever." (Argument from career impact is not valid, but I say it to leave a line of retreat<sup>12</sup>.)

Chalmers critiques substance dualism on the grounds that it's hard to see what new theory of physics, what new substance that interacts with matter, could possibly explain consciousness. But property dualism has exactly the same problem. No matter what kind of dual property you talk about, how exactly does it explain consciousness?

When Chalmers postulated an extra property that *is* consciousness, he *took* that leap across the unexplainable. How does it help his theory to further specify that this extra property *has no effect*? Why not just let it be causal?

<sup>10</sup>. Page 154, 'Mysterious Answers to Mysterious Questions'.

<sup>11</sup>. Page 147, 'Fake Causality'.

<sup>12</sup>. Page 653, 'Leave a Line of Retreat'.

If I were going to be unkind, this would be the time to drag in the dragon—to mention Carl Sagan’s parable of the dragon in the garage<sup>13</sup>. “I have a dragon in my garage.” Great! I want to see it, let’s go! “You can’t see it—it’s an invisible dragon.” Oh, I’d like to hear it then. “Sorry, it’s an inaudible dragon.” I’d like to measure its carbon dioxide output. “It doesn’t breathe.” I’ll toss a bag of flour into the air, to outline its form. “The dragon is permeable to flour.”

One motive for trying to make your theory unfalsifiable, is that deep down you fear to put it to the test. Sir Roger Penrose (physicist) and Stuart Hameroff (neurologist) are substance dualists; they think that there is something mysterious going on in quantum, that Everett is wrong and that the “collapse of the wave-function” is physically real, and that this is where consciousness lives and how it exerts causal effect upon your lips when you say aloud “I think therefore I am.” Believing this, they predicted that neurons would protect themselves from decoherence long enough to maintain macroscopic quantum states.

This is in the process of being tested, and so far, prospects are not looking good for Penrose—

—but Penrose’s basic conduct is scientifically respectable. Not Bayesian, maybe<sup>14</sup>, but still fundamentally healthy. He came up with a wacky hypothesis. He said how to test it. He went out and tried to actually test it.

As I once said to Stuart Hameroff, “I think the hypothesis you’re testing is completely hopeless, and your experiments should *definitely* be funded. Even if you don’t find exactly what you’re looking for, you’re looking in a place where no one else is looking, and you might find something interesting.”

So a nasty dismissal of epiphenomenalism would be that zombie-ists are afraid to say the consciousness-stuff can have *effects*, because then scientists could go *looking* for the extra properties, and fail to find them.

I don’t think this is actually true of Chalmers, though. If Chalmers lacked self-honesty, he could make things a *lot* easier on himself.

---

13. Page 108, ‘Belief in Belief’.

14. Page 167, ‘My Wild and Reckless Youth’.

(But just in case Chalmers is reading this and does have falsification-fear, I'll point out that if epiphenomenalism is false, then there *is* some other explanation for that-which-we-call consciousness, and it will eventually be found, leaving Chalmers's theory in ruins; so if Chalmers cares about his place in history, he has no motive to endorse epiphenomenalism unless he really thinks it's true<sup>15</sup>.)

Chalmers is one of the most frustrating philosophers I know. Sometimes I wonder if he's pulling an "Atheism Conquered"<sup>16</sup>. Chalmers does this really *sharp* analysis... and then turns left at the last minute. He lays out everything that's wrong with the Zombie World scenario, and then, having reduced the whole argument to smithereens, calmly accepts it.

Chalmers does the same thing when he lays out, in calm detail, the problem with saying that our own beliefs in consciousness are justified, when our zombie twins say exactly the same thing for exactly the same reasons and are wrong.

On Chalmers's theory, Chalmers saying that he believes in consciousness cannot be *causally* justified; the belief is not caused by the fact itself<sup>17</sup>. In the absence of consciousness, Chalmers would write the same papers for the same reasons.

On epiphenomenalism, Chalmers saying that he believes in consciousness cannot be justified as the product of a process that systematically outputs true beliefs, because the zombie twin writes the same papers using the same systematic process and is wrong.

Chalmers admits this. Chalmers, in fact, explains the argument in great detail in his book. Okay, so Chalmers has solidly proven that he is not justified in believing in epiphenomenal consciousness, right? No. Chalmers writes:

Conscious experience lies at the center of our epistemic universe; we have access to it *directly*. This raises the question: what is it that justifies our beliefs about our experiences, if it is not a causal link to those experiences, and if it is not the mechanisms by which the beliefs are formed? I think the answer to this is clear: it is *having*

---

<sup>15</sup>. Page 640, 'The Meditation on Curiosity'.

<sup>16</sup>. <http://www.expatsingapore.com/forum/index.php?topic=16763.msg126809>

<sup>17</sup>. Page 703, 'Explaining vs. Explaining Away'.

the experiences that justifies the beliefs. For example, the very fact that I have a red experience now provides justification for my belief that I am having a red experience...

Because my zombie twin lacks experiences, he is in a very different epistemic situation from me, and his judgments lack the corresponding justification. It may be tempting to object that if my belief lies in the physical realm, its justification must lie in the physical realm; but this is a *non sequitur*. From the fact that there is no justification in the physical realm, one might conclude that the *physical* portion of me (my brain, say) is not justified in its belief. But the question is whether *I* am justified in the belief, not whether my *brain* is justified in the belief, and if property dualism is correct than there is more to me than my brain.

So—if I've got this thesis right—there's a core you, above and beyond your brain, that believes it is not a zombie, and directly experiences not being a zombie; and so its beliefs are justified.

But Chalmers just *wrote all that stuff down*, in his very physical book, and so did the zombie-Chalmers.

The zombie Chalmers can't have written the book *because* of the zombie's core self above the brain; there must be some entirely different reason, within the laws of physics.

It follows that even if there *is* a part of Chalmers hidden away that is conscious and believes in consciousness, directly and without mediation, there is also a *separable subspace* of Chalmers—a causally closed cognitive subsystem that acts entirely *within* physics—and this “outer self” is what speaks Chalmers's internal narrative, and writes papers on consciousness.

I do not see any way to evade the charge that, on Chalmers's own theory, this separable outer Chalmers is deranged. This is the part of Chalmers that is the same in this world, or the Zombie World; and in either world it writes philosophy papers on consciousness *for no valid reason*. Chalmers's philosophy papers are not output by that inner core of awareness and belief-in-awareness, they

are output by the mere physics of the internal narrative that makes Chalmers's fingers strike the keys of his computer.

And yet this deranged outer Chalmers is writing philosophy papers that *just happen* to be perfectly right<sup>18</sup>, *by a separate and additional miracle*. Not a logically necessary miracle (then the Zombie World would not be logically possible). A physically contingent miracle, that happens to be true in what we think is our universe, even though science can never distinguish our universe from the Zombie World.

Or at least, that would seem to be the implication of what the self-confessedly deranged outer Chalmers is telling us.

I think I speak for all reductionists when I say *Huh?*

That's not epicycles. That's, "Planetary motions follow these epicycles—but epicycles don't actually *do* anything—there's something else that makes the planets move the same way the epicycles say they should, which I haven't been able to explain—and by the way, I would say this even if there weren't any epicycles."

I have a nonstandard perspective<sup>19</sup> on philosophy because I look at everything with an eye to designing an AI; specifically, a self-improving Artificial General Intelligence with stable motivational structure.

When I think about designing an AI, I ponder principles like probability theory<sup>20</sup>, the Bayesian<sup>21</sup> notion of evidence as differential diagnostic<sup>22</sup>, and above all, reflective coherence. Any self-modifying AI that starts out in a reflectively inconsistent state won't stay that way for long<sup>23</sup>.

If a self-modifying AI looks at a part of itself that concludes "B" on condition A—a part of itself that writes "B" to memory whenever condition A is true—and the AI inspects this part, determines how it (causally) operates in the context of the larger universe, and the AI decides that this part systematically tends to write false data to memory, then the AI has found what appears to be a bug, and the

18. [http://lesswrong.com/lw/06/perpetual\\_motion\\_beliefs/](http://lesswrong.com/lw/06/perpetual_motion_beliefs/)

19. Page 465, 'How to Seem (and Be) Deep'.

20. [http://lesswrong.com/lw/mt/beautiful\\_probability/](http://lesswrong.com/lw/mt/beautiful_probability/)

21. <http://yudowsky.net/bayes/bayes.html>

22. Page 86, 'What is Evidence?'.

23. [http://lesswrong.com/lw/nc/newcombs\\_problem\\_and\\_regret\\_of\\_rationality/](http://lesswrong.com/lw/nc/newcombs_problem_and_regret_of_rationality/)

AI will self-modify not to write “B” to the belief pool under condition A.

Any epistemological theory that disregards reflective coherence is not a good theory to use in constructing self-improving AI. This is a knockdown argument from my perspective, considering what I intend to actually use philosophy *for*. So I have to invent a reflectively coherent theory anyway. And when I do, by golly, reflective coherence turns out to make intuitive sense<sup>24</sup>.

So that’s the unusual way in which I tend to think about these things. And now I look back at Chalmers:

The causally closed “outer Chalmers” (that is not influenced in any way by the “inner Chalmers” that has separate additional awareness and beliefs) must be carrying out some systematically unreliable, unwarranted operation which *in some unexplained fashion* causes the internal narrative to produce beliefs about an “inner Chalmers” that are *correct for no logical reason* in what happens to be our universe.

But there’s no possible warrant for the outer Chalmers *or any reflectively coherent self-inspecting AI* to believe in this mysterious correctness. A good AI design should, I think, look like a reflectively coherent intelligence embodied in a causal system, with a *testable* theory of how that selfsame causal system produces systematically accurate<sup>25</sup> beliefs on the way to achieving its goals<sup>26</sup>.

So the AI will scan Chalmers and see a closed causal cognitive system producing an internal narrative that is uttering nonsense. Nonsense that seems to have a high impact on what Chalmers thinks *should be considered a morally valuable person*.

This is not a *necessary* problem for Friendly AI theorists. It is *only* a problem if you happen to be an epiphenomenalist. If you believe either the reductionists (consciousness happens *within* the atoms) or the substance dualists (consciousness is *causally potent* immaterial stuff), people talking about consciousness are talking about something real, and a reflectively consistent Bayesian AI can see this by tracing back the chain of causality for what makes people say “consciousness”.

---

<sup>24</sup>. Page 99, ‘The Lens That Sees Its Flaws’.

<sup>25</sup>. Page 695, ‘Qualitatively Confused’.

<sup>26</sup>. Page 644, ‘Something to Protect’.

According to Chalmers, the causally closed cognitive system of Chalmers's internal narrative is (mysteriously) malfunctioning in a way that, not by necessity, but just in *our* universe, miraculously happens to be correct. Furthermore, the internal narrative asserts "the internal narrative is mysteriously malfunctioning, but miraculously happens to be correctly echoing the justified thoughts of the epiphenomenal inner core", and again, in *our* universe, miraculously happens to be correct.

*Oh, come on!*

Shouldn't there come a point where you just give up on an idea? Where, on some raw intuitive level, you just go: *What on Earth was I thinking?*

Humanity has accumulated some broad experience with what correct theories of the world look like. *This is not what a correct theory looks like.*

"Argument from incredulity," you say. Fine, you want it spelled out? The said Chalmersian theory postulates multiple unexplained complex miracles. This drives down its prior probability, by the conjunction rule of probability<sup>27</sup> and Occam's Razor<sup>28</sup>. It is therefore dominated by at least two theories which postulate fewer miracles, namely:

- Substance dualism:
  - There is a stuff of consciousness which is not yet understood, an extraordinary super-physical stuff that *visibly affects* our world; and this stuff is what makes us talk about consciousness.
- Not-quite-faith-based<sup>29</sup> reductionism:
  - That-which-we-name<sup>30</sup> "consciousness" happens *within* physics, in a way not yet understood, just like what happened the last three thousand times humanity ran into something mysterious.
  - Your intuition that no material substance can possibly add up<sup>31</sup> to consciousness is incorrect. If you *actually* knew *exactly* why you talk about

---

27. [http://lesswrong.com/lw/jk/burdensome\\_details/](http://lesswrong.com/lw/jk/burdensome_details/)

28. Page 95, 'Occam's Razor'.

29. Page 599, 'The Fallacy of Gray'.

30. Page 766, 'Reductive Reference'.

31. Page 679, 'Wrong Questions'.

consciousness, this would give you new insights, of a form you can't now anticipate; and afterward you would realize that your arguments about normal physics having no room for consciousness were flawed<sup>32</sup>.

Compare to:

- Epiphenomenal property dualism:
  - Matter has additional consciousness-properties which are not yet understood. These properties are epiphenomenal with respect to ordinarily observable physics—they make no difference to the motion of particles.
  - *Separately*, there exists a not-yet-understood reason *within normal physics* why philosophers talk about consciousness and invent theories of dual properties.
  - *Miraculously*, when philosophers talk about consciousness, the bridging laws of *our* world are exactly right to make this talk about consciousness correct, even though it arises from a malfunction (drawing of logically unwarranted conclusions) in the causally closed cognitive system that types philosophy papers.

I know I'm speaking from limited experience, here. But based on my limited experience, the Zombie Argument may be a candidate for *the most deranged idea in all of philosophy*.

There are times when, as a rationalist, you have to believe things that seem weird<sup>33</sup> to you. Relativity seems weird, quantum mechanics seems weird, natural selection<sup>34</sup> seems weird.

But these weirdnesses are pinned down by massive evidence. There's a difference between believing something weird because science has confirmed it overwhelmingly—

—versus believing a proposition that seems downright *deranged*, because of a great big complicated philosophical argument centered around unspecified miracles and giant blank spots not even claimed to be understood—

---

<sup>32</sup>. Page 673, 'Dissolving the Question'.

<sup>33</sup>. [http://lesswrong.com/lw/hs/think\\_like\\_reality/](http://lesswrong.com/lw/hs/think_like_reality/)

<sup>34</sup>. [http://lesswrong.com/lw/kr/an\\_alien\\_god/](http://lesswrong.com/lw/kr/an_alien_god/)

—in a case where *even if you accept everything that has been told to you so far*, afterward the phenomenon will still seem like a mystery and still have the same quality of wondrous impenetrability that it had at the start<sup>35</sup>.

The correct thing for a rationalist to say at this point, if all of David Chalmers's arguments seem individually plausible—which they don't seem to me—is:

“Okay... I don't know how consciousness works... I admit that... and maybe I'm approaching the whole problem wrong, or asking the wrong questions... but this zombie business *can't possibly be right*. The arguments aren't nailed down enough to make me believe this—especially when accepting it won't make me feel any less confused. On a core gut level, this just *doesn't look* like the way reality could *really really* work.”

Mind you, I am not saying this is a substitute for careful analytic refutation of Chalmers's thesis. System 1<sup>36</sup> is not a substitute for System 2, though it can help point the way. You still have to track down where the problems are *specifically*.

Chalmers wrote a big book, not all of which is available through free Google preview. I haven't duplicated the long chains of argument where Chalmers lays out the arguments against himself in calm detail. I've just tried to tack on a final refutation of Chalmers's last presented defense, which Chalmers has not yet countered to my knowledge. Hit the ball back into his court, as it were.

But, yes, on a core level, the *sane* thing to do when you see the conclusion of the zombie argument, is to say “That *can't possibly be right*” and start looking for a flaw.

---

35. Page 154, ‘Mysterious Answers to Mysterious Questions’.

36. Page 83, ‘Why truth? And...’.

## 30. Zombie Responses<sup>1</sup>

**Continuation of:** Zombies! Zombies?<sup>2</sup>

I'm a bit tired today, having stayed up until 3AM writing yesterday's >6000-word post on zombies<sup>3</sup>, so today I'll just reply to Richard, and tie up a loose end I spotted the next day.

Besides, TypePad's nitwit, un-opt-out-able 50-comment pagination "feature", that doesn't work with the Recent Comments sidebar, means that we might as well jump the discussion here before we go over the 50-comment limit.

(A) Richard Chappell writes<sup>4</sup>:

A terminological note (to avoid unnecessary confusion): what you call 'conceivable', others of us would merely call "*apparently conceivable*".

The gap between "I don't see a contradiction yet" and "this is logically possible" is so huge (it's NP-complete even in some simple-seeming cases) that you really should have two different words. As the zombie argument is boosted to the extent that this huge gap can be swept under the rug of minor terminological differences, I really think it would be a good idea to say "conceivable" versus "logically possible" or maybe even have a still more visible distinction. I can't choose professional terminology that has already been established, but in a case like this, I might seriously refuse to use it.

Maybe I will say "apparently conceivable" for the kind of information that zombie advocates get by imagining Zombie Worlds, and "logically possible" for the kind of information that is established by exhibiting a complete model or logical proof. Note the size of the gap between the information you can get by closing your eyes and imagining zombies, and the information you need to carry the argument for epiphenomenalism.

That is, your view would be characterized as a form of Type-A materialism, the view that zombies are not even

---

1. [http://lesswrong.com/lw/p8/zombie\\_responses/](http://lesswrong.com/lw/p8/zombie_responses/)

2. Page 772, 'Zombies! Zombies?'.

3. Page 772, 'Zombies! Zombies?'.

4. [http://www.typepad.com/t/comments?\\_\\_mode=red&user\\_id=475590&id=109409148](http://www.typepad.com/t/comments?__mode=red&user_id=475590&id=109409148)

(genuinely) conceivable, let alone metaphysically possible.

Type-A materialism is a large bundle; you shouldn't attribute the bundle to me until you see me agree with each of the parts. I think that someone who asks "What is consciousness?" is asking a legitimate question, has a legitimate demand for insight; I don't necessarily think that the *answer* takes the form of "Here is this stuff that has all the properties you would attribute to consciousness, for such-and-such reason", but may to some extent consist of insights that cause you to realize you were asking the question the wrong way.

This is not being eliminative about consciousness. It is being realistic about what kind of insights to expect, faced with a problem that (1) seems like it must have *some* solution, (2) seems like it cannot possibly have any solutions<sup>5</sup>, and (3) is being *discussed* in a fashion that has a great big dependence on the not-fully-understood ad-hoc architecture of human cognition.

(1) You haven't, so far as I can tell, identified any *logical contradiction* in the description of the zombie world.

You've just pointed out that it's kind of strange. But there are many bizarre possible worlds out there. That's no reason to posit an implicit contradiction. So it's still completely mysterious to me what this alleged contradiction is supposed to be.

Okay, I'll spell it out from a materialist standpoint:

1. The zombie world, by definition, contains all parts of our world that are closed with respect to causality. In particular, it contains the cause of my saying, "I think therefore I am."
2. When I focus my inward awareness on my inward awareness, I shortly thereafter experience my internal narrative saying "I am focusing my inward awareness on my inward awareness", and can, if I choose, say so out loud.

---

5. Page 679, 'Wrong Questions'.

3. Intuitively, it sure seems like my inward awareness is causing my internal narrative to say certain things.
4. The word “consciousness”, if it has any meaning at all, refers<sup>6</sup> to that-which-is or that-which-causes or that-which-makes-me-say-I-have inward awareness.
5. From (3) and (4) it would follow that if the zombie world is closed with respect to the causes of my saying “I think therefore I am”, the zombie world contains that which we refer to as “consciousness”.
6. By definition, the zombie world does not contain consciousness.
7. (3) seems to me to have a rather high probability of being empirically true. Therefore I evaluate a high empirical probability that the zombie world is logically impossible.

You can save the Zombie World by letting the cause of my internal narrative’s saying “I think therefore I am” be something entirely other than consciousness. In conjunction with the assumption that consciousness does exist, this is the part that struck me as deranged.

But if the above is *conceivable*, then isn’t the Zombie World conceivable?

No, because the two constructions of the Zombie World involve giving the word “consciousness” different empirical referents, like “water” in our world meaning H<sub>2</sub>O versus “water” in Putnam’s Twin Earth meaning XYZ. For the Zombie World to be logically possible, it does not suffice that, for all *you* knew about how the empirical world worked, the word “consciousness” *could* have referred<sup>7</sup> to an epiphenomenon that is entirely different from the consciousness we know. The Zombie World lacks consciousness, not “consciousness”—it is a world without H<sub>2</sub>O, not a world without “water”. This is what is required to carry the empirical statement, “You could eliminate the referent of whatever is meant by “consciousness” from our world, while keeping all the atoms in the same place.”

Which is to say: I hold that it is an *empirical* fact, given what the word “consciousness” actually refers to<sup>8</sup>, that it is *logically* impossi-

---

6. Page 766, ‘Reductive Reference’.

7. Page 766, ‘Reductive Reference’.

8. Page 766, ‘Reductive Reference’.

ble to eliminate consciousness without moving any atoms. What it would mean to eliminate “consciousness” from a world, rather than consciousness, I will not speculate.

(2) It’s misleading to say it’s “miraculous” (on the property dualist view) that our qualia line up so neatly with the physical world. There’s a natural law which guarantees this, after all. So it’s no more miraculous than any other logically contingent nomic necessity (e.g. the constants in our physical laws).

It is the natural law itself that is “miraculous”—counts as an additional complex-improbable element of the theory to be postulated, without having been itself justified in terms of things already known. One postulates (a) an inner world that is conscious (b) a malfunctioning outer world that talks about consciousness for no reason (c) that the two align perfectly. C does not follow from A and B, and so is a separate postulate.

I agree that this usage of “miraculous” conflicts with the philosophical sense of violating a natural law; I meant it in the sense of improbability appearing from no apparent source, a la perpetual motion belief<sup>9</sup>. Hence the word was ill-chosen in context.

That is, Zombie (or ‘Outer’) Chalmers doesn’t actually conclude *anything*, because his utterances are meaningless. A fortiori, he doesn’t conclude anything unwarrantedly. He’s just making noises; these are no more susceptible to epistemic assessment than the chirps of a bird.

Looking at this from an AI-design standpoint, it seems to me like you should be able to build an AI that systematically refines an inner part of itself that correlates (in the sense of mutual information or systematic relations) to the environment, perhaps including floating-point numbers of a sort that I would call “probabilities” because they obey the internal relations mandated by Cox’s Theorems when the AI encounters new information—pardon me, new sense inputs.

---

9. [http://lesswrong.com/lw/o6/perpetual\\_motion\\_beliefs/](http://lesswrong.com/lw/o6/perpetual_motion_beliefs/)

You will say that, unless the AI is more than mere transistors—unless it has the dual aspect—the AI has no beliefs.

I think my views on this were expressed pretty clearly in “The Simple Truth<sup>10</sup>“.

To me, it seems pretty straightforward to construct notions of maps that correlate to territories in systematic ways, without mentioning anything other than things of pure physical causality. The AI outputs a map of Texas. Another AI flies with the map to Texas and checks to see if the highways are in the corresponding places, chirping “True” when it detects a match and “False” when it detects a mismatch. You can refuse to call this “a map of Texas” but the AIs themselves are still chirping “True” or “False”, and the said AIs are going to chirp “False” when they look at Chalmers’s belief in an epiphenomenal inner core, and I for one would agree with them.

It’s clear that the *function of mapping reality* is performed strictly by Outer Chalmers. The whole business of *producing belief representations* is handled by Bayesian structure<sup>11</sup> in causal interactions<sup>12</sup>. There’s nothing left for the Inner Chalmers to do, but bless the whole affair with epiphenomenal *meaning*. So when it comes to talking about “accuracy”, let alone “systematic accuracy”, it seems like we should be able to determine it strictly by looking at the Outer Chalmers.

**(B)** In yesterday’s text, I left out an assumption when I wrote:

If a self-modifying AI looks at a part of itself that concludes “B” on condition A—a part of itself that writes “B” to memory whenever condition A is true—and the AI inspects this part, determines how it (causally) operates in the context of the larger universe, and the AI decides that this part systematically tends to write false data to memory, then the AI has found what appears to be a bug, and the AI will self-modify not to write “B” to the belief pool under condition A...

But there’s no possible warrant for the outer Chalmers or any reflectively coherent self-inspecting AI to believe in this

<sup>10</sup>. <http://yudkowsky.net/bayes/truth.html>

<sup>11</sup>. [http://lesswrong.com/lw/07/searching\\_for\\_bayesstructure/](http://lesswrong.com/lw/07/searching_for_bayesstructure/)

<sup>12</sup>. Page 86, ‘What is Evidence?’.

mysterious correctness. A good AI design should, I think, look like a reflectively coherent intelligence embodied in a causal system, with a *testable* theory of how that selfsame causal system produces systematically accurate<sup>13</sup> beliefs on the way to achieving its goals<sup>14</sup>.

Actually, you need an additional assumption to the above, which is that a “good AI design” (the kind I was thinking of, anyway) judges its own rationality in a modular way; it enforces global rationality by enforcing local rationality. If there is a piece that, relative to its context, is locally systematically unreliable—for some possible beliefs “B<sub>i</sub>” and conditions A<sub>i</sub>, it adds some “B<sub>i</sub>” to the belief pool under local condition A<sub>i</sub>, where reflection by the system indicates that B<sub>i</sub> is not true (or in the case of probabilistic beliefs, not accurate) when the local condition A<sub>i</sub> is true, then this is a bug.

This kind of modularity is a way to make the problem tractable, and it’s how I currently think about the first-generation AI design, but it may not be the only way to make the problem tractable. Obviously a lot of handwaving here, but you get the general idea.

The notion is that a causally closed cognitive system—such as an AI designed by its programmers to use only causally efficacious parts, or an AI whose theory of its own functioning is entirely testable, or the outer Chalmers that writes philosophy papers—which believes that it has an epiphenomenal inner self, must be doing something systematically unreliable because it would conclude the same thing in a Zombie World. A mind all of whose parts are systematically locally reliable, relative to their contexts, would be systematically globally reliable. Ergo, a mind which is globally unreliable must contain at least one locally unreliable part. So a causally closed cognitive system inspecting itself for local reliability must discover that at least one step involved in adding the belief of an epiphenomenal inner self, is unreliable.

If there are other ways for minds to be reflectively coherent which avoid this proof of disbelief in zombies, philosophers are welcome to try and specify them.

The reason why I have to specify all this is that otherwise you get a kind of extremely cheap reflective coherence where the AI can

---

<sup>13</sup>. Page 695, ‘Qualitatively Confused’.

<sup>14</sup>. Page 644, ‘Something to Protect’.

never label itself unreliable. E.g. if the AI finds a part of itself that computes  $2 + 2 = 5$  (in the surrounding context of counting sheep) the AI will reason: “Well, this part malfunctions and says that  $2 + 2 = 5$ ... but by pure coincidence,  $2 + 2$  *is* equal to 5, or so it seems to me... so while the part looks systematically unreliable, I better keep it the way it is, or it will handle this special case wrong.” That’s why I talk about enforcing global reliability by enforcing local systematic reliability—if you just compare your global beliefs to your global beliefs, you don’t go anywhere.

This does have a general lesson: Show your arguments are globally reliable by virtue of each step being locally reliable, don’t just compare the arguments’ conclusions to your intuitions.

(C) An anonymous poster wrote:

A sidepoint, this, but I believe your etymology for “n’shama” is wrong. It is related to the word for “breath”, not “hear”. The root for “hear” contains an ayin, which n’shama does not.

Now that’s what I call a miraculously misleading coincidence—although the word N’Shama arose for completely different reasons, it sounded *exactly the right way* to make me think it referred to an inner listener.

Oops.

## 31. The Generalized Anti-Zombie Principle<sup>1</sup>

**Followup to:** Zombies! Zombies?<sup>2</sup>

“Each problem that I solved became a rule which served afterwards to solve other problems.”

—Rene Descartes, *Discours de la Methode*

“Zombies<sup>3</sup>” are putatively beings that are atom-by-atom identical to us, governed by all the same third-party-visible physical laws, except that they are not conscious.

Though the philosophy is complicated, the core argument against zombies<sup>4</sup> is simple: When you focus your inward awareness on your inward awareness, soon after your internal narrative (the little voice inside your head that speaks your thoughts) says “I am aware of being aware”, and then you say it out loud, and then you type it into a computer keyboard, and create a third-party visible blog post.

Consciousness, whatever it may be—a substance, a process, a name for a confusion—is not epiphenomenal; your mind can catch the inner listener in the act of listening, and say so out loud. *The fact that I have typed this paragraph* would at least *seem* to refute the idea that consciousness has no experimentally detectable consequences.

I hate to say “So now let’s accept this and move on,” over such a philosophically controversial question, but it seems like a considerable majority of Overcoming Bias commenters do accept this. And there are other conclusions you can only get to after you accept that you cannot subtract consciousness and leave the universe looking exactly the same. So now let’s accept this and move on.

The form of the Anti-Zombie Argument seems like it should generalize, becoming an Anti-Zombie Principle. But what is the proper generalization?

Let’s say, for example, that someone says: “I have a switch in my hand, which does not affect your brain in any way; and iff this switch is flipped, you will cease to be conscious.” Does the Anti-

---

1. [http://lesswrong.com/lw/p9/the\\_generalized\\_antizombie\\_principle/](http://lesswrong.com/lw/p9/the_generalized_antizombie_principle/)

2. Page 772, ‘Zombies! Zombies?’.

3. Page 772, ‘Zombies! Zombies?’.

4. Page 772, ‘Zombies! Zombies?’.

Zombie Principle rule this out as well, with the same structure of argument?

It appears to me that in the case above, the answer is yes. In particular, you can say: “Even after your switch is flipped, I will still talk about consciousness *for exactly the same reasons* I did before. If I am conscious right now, I will still be conscious after you flip the switch.”

Philosophers may object, “But now you’re equating consciousness with talking about consciousness! What about the Zombie Master, the chatbot that regurgitates a remixed corpus of amateur human discourse on consciousness?”

But I did *not* equate “consciousness” with verbal behavior. The core premise is that, *among other things*, the true referent<sup>5</sup> of “consciousness” is *also* the *cause in humans* of talking about inner listeners.

As I argued (at some length) in the sequence on words<sup>6</sup>, what you want in defining a word is not always a perfect Aristotelian<sup>7</sup> necessary-and-sufficient definition; sometimes you just want a treasure map<sup>8</sup> that leads you to the extensional referent. So “that which *does in fact* make me talk about an unspeakable awareness” is not a necessary-and-sufficient definition. But if what does *in fact* cause me to discourse about an unspeakable awareness, is not “consciousness”, then...

...then the discourse gets pretty futile. That is not a knockdown argument against zombies—an empirical<sup>9</sup> question can’t be settled by mere difficulties of discourse. But if you try to defy the Anti-Zombie Principle, you will have problems with the *meaning* of your discourse, not just its plausibility.

Could we *define* the word “consciousness” to mean “whatever actually makes humans talk about ‘consciousness’”? This would have the powerful advantage of guaranteeing that there is at least one real fact named by the word “consciousness”. Even if our belief in consciousness is a confusion, “consciousness” would name the cognitive architecture that generated the confusion. But to establish a definition is only to promise to use a word consistently; it doesn’t

---

5. Page 766, ‘Reductive Reference’.

6. Page 301, ‘37 Ways That Words Can Be Wrong’.

7. Page 193, ‘The Parable of Hemlock’.

8. Page 200, ‘Extensions and Intentions’.

9. Page 191, ‘The Parable of the Dagger’.

settle any empirical questions, such as whether our inner awareness makes us talk about our inner awareness.

Let's return to the Off-Switch.

If we allow that the Anti-Zombie Argument applies against the Off-Switch, then the Generalized Anti-Zombie Principle does *not* say only, “Any change that is not in-principle experimentally detectable (IPED) cannot remove your consciousness.” The switch’s flipping is experimentally detectable, but it still seems *highly* unlikely to remove your consciousness.

Perhaps the Anti-Zombie Principle says, “Any change that does not affect you in any IPED way cannot remove your consciousness”?

But is it a reasonable stipulation to say that flipping the switch does not affect you in *any* IPED way? All the particles in the switch are interacting with the particles composing your body and brain. There are gravitational effects—tiny, but real and IPED. The gravitational pull from a one-gram switch ten meters away is around<sup>10</sup>  $6 \times 10^{-16}$  m/s<sup>2</sup>. That’s around half a neutron diameter per second per second, far below thermal noise, but way above the Planck level.

We could flip the switch light-years away, in which case the flip would have no immediate causal effect on you (whatever “immediate” means in this case) (if the Standard Model of physics is correct).

But it doesn’t seem like we *should* have to alter the thought experiment in this fashion. It seems that, if a disconnected switch is flipped on the other side of a room, you should not expect your inner listener to go out like a light, because the switch “obviously doesn’t change” that which is the true cause of your talking about an inner listener. Whatever you really are, you don’t expect the switch to mess with it.

This is a *large* step.

If you deny that it is a reasonable step, you had better never go near a switch again. But still, it’s a large step.

The key idea of reductionism<sup>11</sup> is that our maps of the universe are multi-level to save on computing power, but physics seems to

<sup>10.</sup> [http://www.google.com/search?hl=en&safe=off&q=G+\\*+1+gram+2810+meters29%5E2&btnG=Search](http://www.google.com/search?hl=en&safe=off&q=G+*+1+gram+2810+meters29%5E2&btnG=Search)

<sup>11.</sup> Page 699, ‘Reductionism’.

be strictly single-level. All our discourse about the universe takes place using references far above<sup>12</sup> the level of fundamental particles.

The switch's flip *does* change the fundamental particles of your body and brain. It nudges them by whole neutron diameters away from where they would have otherwise been.

In ordinary life, we gloss a change this small by saying that the switch "doesn't affect you". But it *does* affect you. It changes everything by whole neutron diameters! What could possibly be remaining the same? Only the *description* that you would give of the higher levels of organization—the cells, the proteins, the spikes traveling along a neural axon. As the map is far less detailed than the territory, it must map many different states to the same description<sup>13</sup>.

Any reasonable sort of humanish *description* of the brain that talks about neurons and activity patterns (or even the conformations of individual microtubules making up axons and dendrites) won't change when you flip a switch on the other side of the room. Nuclei are larger than neutrons, atoms are larger than nuclei, and by the time you get up to talking about the *molecular* level, that tiny little gravitational force has vanished from the list of things you bother to *track*.

But if you add up enough tiny little gravitational pulls, they will eventually yank you across the room and tear you apart by tidal forces, so clearly a small effect is *not* "no effect at all".

Maybe the tidal force from that tiny little pull, by an *amazing* coincidence, pulls a single extra calcium ion just a tiny bit closer to an ion channel, causing it to be pulled in just a tiny bit sooner, making a single neuron fire infinitesimally sooner than it would otherwise have done, a difference which amplifies chaotically, finally making a whole neural spike occur that otherwise wouldn't have occurred, sending you off on a different train of thought, that triggers an epileptic fit, that kills you, causing you to cease to be conscious...

If you add up a lot of tiny quantitative effects, you get a big quantitative effect—big enough to mess with anything you care to name. And so claiming that the switch has literally *zero* effect on the things you care about, is taking it too far.

<sup>12</sup>. Page 766, 'Reductive Reference'.

<sup>13</sup>. Page 253, 'Fallacies of Compression'.

But with just one switch, the force exerted is vastly less than thermal uncertainties, never mind quantum uncertainties. If you don't expect your consciousness to flicker in and out of existence as the result of thermal jiggling, then you certainly<sup>14</sup> shouldn't expect to go out like a light when someone sneezes a kilometer away.

The alert Bayesian will note that I have just made an argument about *expectations*, states of *knowledge*, justified *beliefs* about what can and can't switch off your consciousness.

This doesn't necessarily destroy the Anti-Zombie Argument. Probabilities are not certainties, but the *laws of probability* are theorems<sup>15</sup>; if rationality says you can't believe something on your current information, then that is a law, not a suggestion.

Still, this version of the Anti-Zombie Argument is weaker. It doesn't have the nice, clean, absolutely clear-cut status of, "You can't possibly eliminate consciousness while leaving all the atoms in *exactly* the same place." (Or for "all the atoms" substitute "all causes with in-principle experimentally detectable effects", and "same wavefunction" for "same place", etc.)

But the new version of the Anti-Zombie Argument still carries. You can say, "I don't know what consciousness really is, and I suspect I may be fundamentally confused about the question. But if the word refers to anything at all, it refers to something that is, among other things, the cause of my talking about consciousness. Now, I don't know why I talk about consciousness. But it happens inside my skull, and I expect it has something to do with neurons firing. Or maybe, if I really understood consciousness, I would have to talk about an even more fundamental level than that, like microtubules, or neurotransmitters diffusing across a synaptic channel. But still, that switch you just flipped has an effect on my neurotransmitters and microtubules that's much, much less than thermal noise at 310 Kelvin. So whatever the true cause of my talking about consciousness may be, I don't expect it to be hugely affected by the gravitational pull from that switch. Maybe it's just a tiny little infinitesimal bit affected? But it's certainly not going to go out like a light. I expect to go on talking about consciousness in *almost exactly* the same way afterward, for *almost exactly* the same reasons."

---

<sup>14</sup>. Page 604, 'Absolute Authority'.

<sup>15</sup>. [http://lesswrong.com/lw/06/perpetual\\_motion\\_beliefs/](http://lesswrong.com/lw/06/perpetual_motion_beliefs/)

This application of the Anti-Zombie Principle is weaker. But it's also much more general. And, in terms of sheer common sense, correct.

The reductionist and the substance dualist actually have two different versions of the above statement. The reductionist furthermore says, “Whatever makes me talk about consciousness, it seems likely that the important parts take place on a much higher functional level than atomic nuclei. Someone who understood consciousness could abstract away from individual neurons firing, and talk about high-level cognitive architectures, and still describe how my mind produces thoughts like ‘I think therefore I am’. So nudging things around by the diameter of a nucleon, shouldn’t affect my consciousness (except maybe with very small probability, or by a very tiny amount, or not until after a significant delay).”

The substance dualist furthermore says, “Whatever makes me talk about consciousness, it’s got to be something beyond the computational physics we know, which means that it might very well involve quantum effects. But still, my consciousness doesn’t flicker on and off whenever someone sneezes a kilometer away. If it did, I would *notice*. It would be like skipping a few seconds, or coming out of a general anesthetic, or sometimes saying, “I don’t think therefore I’m not.” So since it’s a physical fact that thermal vibrations don’t disturb the stuff of my awareness, I don’t expect flipping the switch to disturb it either.”

Either way, you *shouldn’t* expect your sense of awareness to vanish when someone says the word “Abracadabra”, even if that does have some infinitesimal physical effect on your brain—

But hold on! If you *bear* someone say the word “Abracadabra”, that has a very noticeable effect on your brain—so large, even your brain can notice it. It may alter your internal narrative; you may think, “Why did that person just say ‘Abracadabra’?”

Well, but *still* you expect to go on talking about consciousness in almost exactly the same way afterward, for almost exactly the same reasons.

And again, it’s not that “consciousness” is being *equated* to “that which makes you talk about consciousness”. It’s just that consciousness, *among other things*, makes you talk about consciousness.

So anything that makes your consciousness go out like a light, should make you stop talking about consciousness.

If we do something to you, where you don't see how it could possibly change your internal narrative—the little voice in your head that sometimes says things like “I think therefore I am”, whose words you can choose to say aloud—then it shouldn't make you cease to be conscious.

And this is true even if the internal narrative is just “pretty much the same”, and the causes of it are also pretty much the same; among the causes that are pretty much the same, is whatever you mean by “consciousness”.

If you're wondering where all this is going, and why it's important to go to such tremendous lengths to ponder such an obvious-seeming Generalized Anti-Zombie Principle, then consider the following debate:

Albert: “Suppose I replaced all the neurons in your head with tiny robotic artificial neurons that had the same connections, the same local input-output behavior, and analogous internal state and learning rules.”

Bernice: “That's killing me! There wouldn't be a conscious being there anymore.”

Charles: “Well, there'd still be a conscious being there, but it wouldn't be *me*.”

Sir Roger Penrose: “The thought experiment you propose is impossible. You *can't* duplicate the behavior of neurons without tapping into quantum gravity. That said, there's not much point in me taking further part in this conversation.” (*Wanders away*)

Albert: “Suppose that the replacement is carried out one neuron at a time, and the swap occurs so fast that it doesn't make any difference to global processing.”

Bernice: “How could that possibly be the case?”

Albert: “The little robot swims up to the neuron, surrounds it, scans it, learns to duplicate it, and then suddenly takes over the behavior, between one spike and the next. In fact, the imitation is *so* good, that your outward behavior is just the same as it would be if the brain were left undisturbed. Maybe not *exactly* the same, but the causal impact is much less than thermal noise at 310 Kelvin.”

Charles: “So what?”

Albert: “So don’t your beliefs violate the Generalized Anti-Zombie Principle? Whatever just happened, it didn’t change your internal narrative! You’ll go around talking about consciousness for exactly the same reason as before.”

Bernice: “Those little robots are a Zombie Master. They’ll make me talk about consciousness even though I’m not conscious. The Zombie World is possible if you allow there to be an added, extra, experimentally detectable Zombie Master—which those robots are.”

Charles: “Oh, that’s not right, Bernice. The little robots aren’t plotting how to fake consciousness, or processing a corpus of text from human amateurs. They’re doing the same thing neurons do, just in silicon instead of carbon.”

Albert: “Wait, didn’t you just agree with me?”

Charles: “I never said the new person wouldn’t be conscious. I said it wouldn’t be *me*.”

Albert: “Well, obviously the Anti-Zombie Principle generalizes to say that this operation hasn’t disturbed the true cause of your talking about this *me* thing.”

Charles: “Uh-uh! Your operation certainly did disturb the true cause of my talking about consciousness. It substituted a *different* cause in its place, the robots. Now, just because that new cause *also* happens to be conscious—talks about consciousness for the same *generalized* reason—doesn’t mean it’s the *same* cause that was originally there.”

Albert: “But I wouldn’t even have to *tell* you about the robot operation. You wouldn’t *notice*. If you think, going on introspective evidence, that you are in an important sense “the same person” that you were five minutes ago, and I do something to you that doesn’t change the introspective evidence available to you, then your conclusion that you are the same person that you were five minutes ago should be equally justified. Doesn’t the Generalized Anti-Zombie Principle say that if I do something to you that alters your consciousness, let alone makes you a completely different person, then you ought to *notice* somehow?”

Bernice: “Not if you replace me with a Zombie Master. Then there’s no one there *to notice*.”

Charles: “Introspection isn’t perfect. Lots of stuff goes on inside my brain that I don’t notice.”

Albert: “You’re postulating epiphenomenal facts about consciousness and identity!”

Bernice: “No I’m not! I can experimentally detect the difference between neurons and robots.”

Charles: “No I’m not! I can experimentally detect the moment when the old me is replaced by a new person.”

Albert: “Yeah, and I can detect the switch flipping! You’re detecting something that doesn’t *make a noticeable difference* to the *true cause* of your talk about consciousness and personal identity. And the proof is, you’ll talk just the same way afterward.”

Bernice: “That’s because of your robotic Zombie Master!”

Charles: “Just because two people talk about ‘personal identity’ for similar reasons doesn’t make them the same person.”

I think the Generalized Anti-Zombie Principle supports Albert’s position, but the reasons shall have to wait for future posts. I need other prerequisites, and besides, this post is already too long.

But you see the importance of the question, “How far can you generalize the Anti-Zombie Argument<sup>16</sup> and have it still be valid?”

The makeup of future galactic civilizations may be determined by the answer...

---

<sup>16</sup>16. Page 772, ‘Zombies! Zombies?’.

## 32. GAZP vs. GLUT<sup>1</sup>

**Followup to:** The Generalized Anti-Zombie Principle<sup>2</sup>

In “The Unimagined Preposterousness of Zombies”, Daniel Dennett says:

To date, several philosophers have told me that they plan to accept my challenge to offer a non-question-begging defense of zombies, but the only one I have seen so far involves postulating a “logically possible” but fantastic being — a descendent of Ned Block’s Giant Lookup Table fantasy...

A Giant Lookup Table, in programmer’s parlance, is when you implement a function as a giant table of inputs and outputs, usually to save on runtime computation. If my program needs to know the multiplicative product of two inputs between 1 and 100, I can write a multiplication algorithm that computes each time the function is called, or I can precompute a Giant Lookup Table with 10,000 entries and two indices. There are times when you *do* want to do this, though not for multiplication—times when you’re going to reuse the function a lot and it doesn’t have many possible inputs; or when clock cycles are cheap while you’re initializing, but very expensive while executing.

Giant Lookup Tables get very large, very fast. A GLUT of all possible twenty-ply conversations with ten words per remark, using only 850-word Basic English, would require  $7.6 * 10^{85}$  entries.

Replacing a human brain with a Giant Lookup Table of all possible sense inputs and motor outputs (relative to some fine-grained digitization scheme) would require an *unreasonably large amount* of memory storage. But “in principle”, as philosophers are fond of saying, it could be done.

The GLUT is not a zombie in the classic sense, because it is microphysically dissimilar to a human. (In fact, a GLUT can’t *really* run on the same physics as a human; it’s too large to fit in our uni-

---

1. [http://lesswrong.com/lw/pa/gazp\\_vs\\_glut/](http://lesswrong.com/lw/pa/gazp_vs_glut/)

2. Page 799, ‘The Generalized Anti-Zombie Principle’.

verse. For philosophical purposes, we shall ignore this and suppose a supply of unlimited memory storage.)

But is the GLUT a zombie at *all*? That is, does it behave exactly like a human without being conscious?

The GLUT-ed body's tongue talks about consciousness. Its fingers write philosophy papers. In every way, so long as you don't peer inside the skull, the GLUT seems just like a human... which certainly seems like a valid example of a zombie: it behaves just like a human, but there's no one home.

Unless the GLUT is conscious, in which case it wouldn't be a valid example.

I can't recall ever seeing *anyone* claim that a GLUT is conscious. (Admittedly my reading in this area is not up to professional grade; feel free to correct me.) Even people who are accused of being (gasp!) functionalists don't claim that GLUTs can be conscious.

GLUTs are the *reductio ad absurdum* to anyone who suggests that consciousness *is simply* an input-output pattern, thereby disposing of all troublesome worries about what goes on inside.

So what does the Generalized Anti-Zombie Principle<sup>3</sup> (GAZP) say about the Giant Lookup Table (GLUT)?

At first glance, it would seem that a GLUT is the very archetype of a Zombie Master—a distinct, additional, detectable, non-conscious system that animates a zombie and makes it talk about consciousness for *different* reasons.

In the interior of the GLUT, there's merely a very simple computer program that looks up inputs and retrieves outputs. Even talking about a "simple computer program" is overshooting the mark, in a case like this. A GLUT is more like ROM than a CPU. We could equally well talk about a series of switched tracks by which some balls roll out of a previously stored stack and into a trough—*period*; that's *all* the GLUT does.

A spokesperson from People for the Ethical Treatment of Zombies replies: "Oh, that's what all the anti-mechanists say, isn't it? That when you look in the brain, you just find a bunch of neurotransmitters opening ion channels? If ion channels can be conscious, why not levers and balls rolling into bins?"

---

3. Page 799, 'The Generalized Anti-Zombie Principle'.

“The problem isn’t the levers,” replies the functionalist, “the problem is that a GLUT has the *wrong pattern* of levers. You need levers that implement things like, say, formation of beliefs about beliefs, or self-modeling... Heck, you need the ability to write things to memory just so that time can pass for the computation. Unless you think it’s possible to program a conscious being in Haskell.”

“I don’t know about that,” says the PETZ spokesperson, “all I know is that this so-called zombie writes philosophical papers about consciousness. Where do these philosophy papers come from, if not from consciousness?”

Good question! Let us ponder it deeply.

There’s a game in physics called Follow-The-Energy. Richard Feynman’s father<sup>4</sup> played it with young Richard:

It was the kind of thing my father would have talked about: “What makes it go? Everything goes because the sun is shining.” And then we would have fun discussing it:

“No, the toy goes because the spring is wound up,” I would say. “How did the spring get wound up?” he would ask.

“I wound it up.”

“And how did you get moving?”

“From eating.”

“And food grows only because the sun is shining. So it’s because the sun is shining that all these things are moving.” That would get the concept across that motion is simply the *transformation* of the sun’s power.

When you get a little older, you learn that energy is conserved, never created or destroyed, so the notion of *using up* energy doesn’t make much sense. You can never change the total amount of energy, so in what sense are you *using* it?

So when physicists grow up, they learn to play a new game called Follow-The-Negentropy<sup>5</sup>—which is really the same game they were playing all along; only the rules are mathier, the game is more useful,

4. <http://www.textbookleague.org/103feyn.htm>

5. [http://lesswrong.com/lw/o5/the\\_second\\_law\\_of\\_thermodynamics\\_and\\_engines\\_of/](http://lesswrong.com/lw/o5/the_second_law_of_thermodynamics_and_engines_of/)

and the principles are harder to wrap your mind around conceptually.

Rationalists learn a game called Follow-The-Improbability<sup>6</sup>, the grownup version of “How Do You Know?” The rule of the rationalist’s game is that every improbable-seeming belief needs an equivalent amount of evidence to justify it.<sup>7</sup> (This game has *amazingly similar* rules to Follow-The-Negentropy.)

Whenever someone violates the rules of the rationalist’s game, you can find a place in their argument where a quantity of improbability appears from nowhere;<sup>8</sup> and this is as much a sign of a problem as, oh, say, an ingenious design of linked wheels and gears that keeps itself running forever.

The one comes to you and says: “I believe with firm and abiding faith that there’s an object in the asteroid belt, one foot across and composed entirely of chocolate cake; you can’t prove<sup>9</sup> that this is impossible.” But, unless the one had access to some kind of evidence for this belief, it would be highly improbable for a correct belief to form *spontaneously*. So either the one can point to evidence, or the belief won’t turn out to be true. “But you can’t prove it’s *impossible* for my mind to spontaneously generate a belief that happens to be correct!” No, but that kind of spontaneous generation is *highly improbable*, just like, oh, say, an egg unscrambling itself.

In Follow-The-Improbability<sup>10</sup>, it’s highly suspicious to even *talk* about a specific hypothesis without having had enough evidence to narrow down the space of possible hypotheses<sup>11</sup>. Why aren’t you giving equal air time to a decillion other equally plausible hypotheses? You need sufficient evidence to find the “chocolate cake in the asteroid belt” hypothesis in the hypothesis space—otherwise there’s no reason to give it more air time than a trillion other candidates like “There’s a wooden dresser in the asteroid belt” or “The Flying Spaghetti Monster threw up on my sneakers.”

---

6. [http://lesswrong.com/lw/06/perpetual\\_motion\\_beliefs/](http://lesswrong.com/lw/06/perpetual_motion_beliefs/)

7. Page 89, ‘How Much Evidence Does It Take?’.

8. [http://lesswrong.com/lw/06/perpetual\\_motion\\_beliefs/](http://lesswrong.com/lw/06/perpetual_motion_beliefs/)

9. Page 604, ‘Absolute Authority’.

10. [http://lesswrong.com/lw/06/perpetual\\_motion\\_beliefs/](http://lesswrong.com/lw/06/perpetual_motion_beliefs/)

11. [http://lesswrong.com/lw/jo/einsteins\\_arrogance/](http://lesswrong.com/lw/jo/einsteins_arrogance/)

In Follow-The-Improbability, you are not allowed to pull out big complicated specific hypotheses from thin air without *already* having a corresponding amount of evidence; because it's not realistic to suppose that you could spontaneously start discussing the *true* hypothesis by *pure coincidence*.

A philosopher says, “This zombie’s skull contains a Giant Lookup Table of all the inputs and outputs for some human’s brain.” This is a very *large* improbability. So you ask, “How did this improbable event occur? Where did the GLUT come from?”

Now this is not standard philosophical procedure for thought experiments. In standard philosophical procedure, you are allowed to postulate things like “Suppose you were riding a beam of light...” without worrying about physical possibility, let alone mere improbability. But in this case, the origin of the GLUT matters; and that’s why it’s important to understand the motivating question, “Where did the improbability come from?”

The obvious answer is that you took a computational specification of a human brain, and used *that* to precompute the Giant Lookup Table. (Thereby creating uncounted googols of human beings, some of them in extreme pain, the supermajority gone quite mad in a universe of chaos where inputs bear no relation to outputs. But damn the ethics, this is for *philosophy*.)

In this case, the GLUT *is* writing papers about consciousness because of a conscious algorithm. The GLUT is no more a zombie, than a cellphone is a zombie because it can talk about consciousness while being just a small consumer electronic device. The cellphone is just transmitting philosophy speeches from whoever happens to be on the other end of the line. A GLUT generated from an originally human brain-specification is doing the same thing.

“All right,” says the philosopher, “the GLUT was generated randomly, and *just happens* to have the same input-output relations as some reference human.”

How, exactly, did you randomly generate the GLUT?

“We used a true randomness source—a quantum device.”

But a quantum device just implements the Branch Both Ways instruction; when you generate a bit from a quantum randomness source, the deterministic result is that one set of universe-branches

(locally connected amplitude clouds) see 1, and another set of universes see 0. Do it 4 times, create 16 (sets of) universes.

So, really, this is like saying that you got the GLUT by writing down all possible GLUT-sized sequences of 0s and 1s, in a really damn huge bin of lookup tables; and then reaching into the bin, and *somewhat* pulling out a GLUT that happened to correspond to a human brain-specification. Where did the improbability come from?

Because if this *wasn't just a coincidence*—if you had some reach-into-the-bin function that pulled out a human-corresponding GLUT by *design*, not just chance—then that reach-into-the-bin function is probably conscious, and so the GLUT is again a cell-phone, not a zombie. It's connected to a human at two removes, instead of one, but it's still a cellphone! Nice try at concealing the source of the improbability there!

Now behold where Follow-The-Improbability has taken us: where is the source of this body's tongue talking about an inner listener? The consciousness isn't in the lookup table. The consciousness isn't in the factory that manufactures lots of possible lookup tables. The consciousness was in whatever *pointed to one particular already-manufactured lookup table*, and said, “Use that one!”

You can see why I introduced the game of Follow-The-Improbability. Ordinarily, when we're talking to a person, we tend to think that whatever is inside the skull, must be “where the consciousness is”. It's only by playing Follow-The-Improbability that we can realize that the real source of the conversation we're having, is that-which-is-responsible-for the *improbability* of the conversation—however distant in time or space, as the Sun moves a wind-up toy.

“No, no!” says the philosopher. “In the thought experiment, they aren't randomly generating lots of GLUTs, and then using a conscious algorithm to pick out one GLUT that seems humanlike! I am *specifying* that, in this thought experiment, they reach into the inconceivably vast GLUT bin, and *by pure chance* pull out a GLUT that is identical to a human brain's inputs and outputs! *There!* I've got you cornered now! You can't play Follow-The-Improbability any further!”

Oh. So your *specification* is the source of the improbability here.

When we play Follow-The-Improbability again, we end up *outside the thought experiment*, looking at the *philosopher*.

That which points to the one GLUT that talks about consciousness, out of all the vast space of possibilities, is now... the conscious person asking us to imagine this whole scenario. And our own brains, which will fill in the blank when we imagine, “What will this GLUT say in response to ‘Talk about your inner listener’?”

The moral of this story is that when you follow back discourse about “consciousness”, you generally find consciousness. It’s not always right in front of you. Sometimes it’s very cleverly hidden. But it’s there. Hence the Generalized Anti-Zombie Principle.

If there is a Zombie Master in the form of a chatbot that processes and remixes amateur human discourse about “consciousness”, the humans who generated the original text corpus are conscious.

If someday you come to understand consciousness, and look back, and see that there’s a program you can write which will output confused philosophical discourse that sounds an awful lot like humans without itself being conscious—then when I ask “How did this program come to sound similar to humans?” the answer is that *you* wrote it to sound similar to *conscious humans*, rather than choosing on the criterion of similarity to something else. This doesn’t mean your little Zombie Master is conscious—but it does mean I can find consciousness somewhere in the universe by tracing back the chain of causality, which means we’re not entirely in the Zombie World.

But suppose someone actually *did* reach into a GLUT-bin and by *genuinely pure chance* pulled out a GLUT that wrote philosophy papers?

Well, then it wouldn’t be conscious. IMHO.

I mean, there’s got to be more to it than inputs and outputs.

Otherwise even a GLUT would be conscious, right?

---

Oh, and for those of you wondering how this sort of thing relates to my day job...

In this line of business you meet an awful lot of people who think that an arbitrarily generated powerful AI will be “moral”. They can’t agree among themselves on why, or what they mean by the word “moral”; but they all agree that doing Friendly AI theory is

unnecessary. And when you ask them how an arbitrarily generated AI ends up with moral outputs, they proffer elaborate rationalizations aimed at AIs<sup>12</sup> of that which they deem “moral”; and there are all sorts of problems with this<sup>13</sup>, but the number one problem is, “Are you *sure* the AI would follow the same line of thought you invented to argue human morals, when, unlike you, the AI doesn’t start out knowing what *you* want it to rationalize?” You could call the counter-principle Follow-The-Decision-Information, or something along those lines. You can account for an AI that does improbably nice things by telling me how you chose the AI’s design from a huge space of possibilities, but otherwise the improbability is being pulled out of nowhere—though more and more heavily disguised, as rationalized premises are rationalized in turn.

So I’ve already done a whole series of posts<sup>14</sup> which I myself generated<sup>15</sup> using Follow-The-Improbability. But I didn’t spell out the rules *explicitly* at that time, because I hadn’t done the thermodynamic<sup>16</sup> posts yet...

Just thought I’d mention that. It’s amazing how many of my Overcoming Bias posts would coincidentally turn out to include ideas surprisingly relevant to discussion of Friendly AI theory... if you believe in coincidence.

---

12. Page 538, ‘Fake Optimization Criteria’.

13. [http://lesswrong.com/lw/lp/fake\\_fake\\_utility\\_functions/](http://lesswrong.com/lw/lp/fake_fake_utility_functions/)

14. [http://lesswrong.com/lw/lp/fake\\_fake\\_utility\\_functions/](http://lesswrong.com/lw/lp/fake_fake_utility_functions/)

15. Page 183, ‘Truly Part Of You’.

16. [http://lesswrong.com/lw/o6/perpetual\\_motion\\_beliefs/](http://lesswrong.com/lw/o6/perpetual_motion_beliefs/)

## 33. Belief in the Implied Invisible<sup>1</sup>

**Followup to:** The Generalized Anti-Zombie Principle<sup>2</sup>

One generalized lesson *not* to learn from the Anti-Zombie Argument is, “Anything you can’t see doesn’t exist.”

It’s tempting to conclude the general rule. It would make the Anti-Zombie Argument much simpler, on future occasions, if we could take this as a premise. But unfortunately that’s just not Bayesian.

Suppose I transmit a photon out toward infinity, not aimed at any stars, or any galaxies, pointing it toward one of the great voids between superclusters. Based on standard physics, in other words, I don’t expect this photon to intercept anything on its way out. The photon is moving at light speed, so I can’t chase after it and capture it again.

If the expansion of the universe is accelerating, as current cosmology holds, there will come a future point where I don’t expect to be able to interact with the photon even in principle—a future time beyond which I don’t expect the photon’s future light cone to intercept my world-line. Even if an alien species captured the photon and rushed back to tell us, they couldn’t travel fast enough to make up for the accelerating expansion of the universe.

Should I believe that, in the moment where I can no longer interact with it even in principle, the photon disappears?

No.

It would violate Conservation of Energy. And the second law of thermodynamics. And just about every other law of physics. And probably the Three Laws of Robotics. It would imply the photon knows I care about it and knows exactly when to disappear.

It’s a *silly idea*.

But if you can believe in the continued existence of photons that have become experimentally undetectable to you, why doesn’t this imply a general license to believe in the invisible?

(If you want to think about this question on your own, do so before the jump...)

---

1. [http://lesswrong.com/lw/pb/belief\\_in\\_the\\_implied\\_invisible/](http://lesswrong.com/lw/pb/belief_in_the_implied_invisible/)

2. Page 799, ‘The Generalized Anti-Zombie Principle’.

Though I failed to Google a source, I remember reading that when it was first proposed that the Milky Way was our *galaxy*—that the hazy river of light in the night sky was made up of millions (or even billions) of stars—that Occam’s Razor was invoked against the new hypothesis. Because, you see, the hypothesis vastly multiplied the number of “entities” in the believed universe. Or maybe it was the suggestion that “nebulae”—those hazy patches seen through a telescope—might be galaxies full of stars, that got the invocation of Occam’s Razor.

*Lex parsimoniae: Entia non sunt multiplicanda praeter necessitatem.*

That was Occam’s original formulation, the law of parsimony: Entities should not be multiplied beyond necessity.

If you postulate billions of stars that no one has ever believed in before, you’re multiplying entities, aren’t you?

No. There are two Bayesian formalizations of Occam’s Razor<sup>3</sup>: Solomonoff Induction, and Minimum Message Length. Neither penalizes galaxies for being big.

Which they had better not do! One of the lessons of history is that what-we-call-reality keeps turning out to be bigger and bigger and huger yet. Remember when the Earth was at the center of the universe? Remember when no one had invented Avogadro’s number? If Occam’s Razor was weighing against the multiplication of entities every time, we’d have to start doubting Occam’s Razor, because it would have consistently turned out to be wrong.

In Solomonoff induction, the complexity of your model is the amount of *code* in the computer program you have to write to simulate your model. The amount of *code*, not the amount of RAM it uses, or the number of cycles it takes to compute. A model of the universe that contains billions of galaxies containing billions of stars, each star made of a billion trillion decillion quarks, will take a lot of RAM to run—but the *code* only has to describe the behavior of the quarks, and the stars and galaxies can be left to run themselves. I am speaking semi-metaphorically here—there are things in the universe besides quarks—but the point is, postulating an extra billion galaxies doesn’t count against the size of your code, if you’ve already described one galaxy. It just takes a bit more RAM, and Occam’s Razor doesn’t care about RAM.

---

3. Page 95, ‘Occam’s Razor’.

Why not? The Minimum Message Length formalism, which is nearly equivalent to Solomonoff Induction, may make the principle clearer: If you have to tell someone how your model of the universe works, you don't have to individually specify the location of each quark in each star in each galaxy. You just have to write down some equations. The amount of "stuff" that obeys the equation doesn't affect how long it takes to write the equation down. If you encode the equation into a file, and the file is 100 bits long, then there are  $2^{100}$  other models that would be around the same file size, and you'll need roughly 100 bits of supporting evidence. You've got a limited amount of probability mass; and a priori, you've got to divide that mass up among all the messages you could send; and so postulating a model from within a model space of  $2^{100}$  alternatives, means you've got to accept a  $2^{-100}$  prior probability penalty—but having more galaxies doesn't add to this.

Postulating billions of stars in billions of galaxies doesn't affect the length of your message describing the overall behavior of all those galaxies. So you don't take a probability hit from having the *same* equations describing more things. (So long as your model's predictive successes aren't sensitive to the exact initial conditions. If you've got to specify the exact positions of all the quarks for your model to predict as well as it does, the extra quarks do count as a hit.)

If you suppose that the photon disappears when you are no longer looking at it, this is an *additional law* in your model of the universe. It's the laws that are "entities", costly under the laws of parsimony. Extra quarks are free.

So does it boil down to, "I believe the photon goes on existing as it wings off to nowhere, because my priors say it's simpler for it to go on existing than to disappear"?

This is what I thought at first, but on reflection, it's not quite right. (And not just because it opens the door to obvious abuses.)

I would boil it down to a distinction between belief in the *implied invisible*, and belief in the *additional invisible*.

When you believe that the photon goes on existing as it wings out to infinity, you're not believing that as an *additional fact*.

What you believe (assign probability to) is a set of simple equations; you believe these equations describe the universe. You be-

lieve these equations because they are the simplest equations you could find that describe the evidence. These equations are *highly* experimentally testable; they explain huge mounds of evidence visible in the past, and predict the results of many observations in the future.

You believe these equations, and it is a *logical implication* of these equations that the photon goes on existing as it wings off to nowhere, so you believe that as well.

Your priors, or even your probabilities, don't *directly* talk about the photon. What you assign probability to is not the photon, but the general laws. When you assign probability to the laws of physics as we know them, you *automatically* contribute that same probability to the photon continuing to exist on its way to nowhere—if you believe the logical implications of what you believe.

It's not that you believe in the invisible *as such*, from reasoning about invisible things. Rather the experimental evidence supports certain laws, and belief in those laws logically implies the existence of certain entities that you can't interact with. This is belief in the *implied invisible*.

On the other hand, if you believe that the photon is eaten out of existence by the Flying Spaghetti Monster—maybe on this just one occasion—or even if you believed without reason that the photon hit a dust speck on its way out—then you would be believing in a specific extra invisible event, on its own. If you thought that this sort of thing happened in general, you would believe in a specific extra invisible law. This is belief in the *additional invisible*.

The whole matter would be a lot simpler, admittedly, if we could just rule out the existence of entities we can't interact with, once and for all—have the universe stop existing at the edge of our telescopes. But this requires us to be very silly.

Saying that you shouldn't ever need a separate and additional belief about invisible things—that you only believe invisibles that are *logical implications* of general laws which are themselves testable, and even then, don't have any further beliefs about them that are not logical implications of visibly testable general rules—actually does seem to rule out all abuses of belief in the invisible, when applied correctly.

Perhaps I should say, “you should assign unaltered prior probability to additional invisibles”, rather than saying, “do not believe in them.” But if you think of a *belief* as something evidentially additional, something you bother to track, something where you bother to count up support for or against, then it’s questionable whether we should ever have additional beliefs about additional invisibles.

There are exotic cases that break this in theory. (E.g: The epiphenomenal demons are watching you, and will torture  $3^{3^3}3^4$  victims for a year, somewhere you can’t ever verify the event, if you ever say the word “Niblick”.) But I can’t think of a case where the principle fails in human practice.

**Added:** To make it clear why you would sometimes want to think about implied invisibles, suppose you’re going to launch a spaceship, at nearly the speed of light, toward a faraway supercluster. By the time the spaceship gets there and sets up a colony, the universe’s expansion will have accelerated too much for them to ever send a message back. Do you deem it worth the purely altruistic effort to set up this colony, for the sake of all the people who will live there and be happy? Or do you think the spaceship blips out of existence before it gets there? This could be a very real question at some point.

---

4. [http://lesswrong.com/lw/kd/pascals\\_mugging\\_tiny\\_probabilities\\_of\\_vast/](http://lesswrong.com/lw/kd/pascals_mugging_tiny_probabilities_of_vast/)

## 34. Zombies: The Movie<sup>1</sup>

FADE IN around a serious-looking group of uniformed military officers. At the head of the table, a senior, heavy-set man, GENERAL FRED, speaks.

GENERAL FRED: The reports are confirmed. New York has been overrun... by *zombies*.

COLONEL TODD: Again? But we just had a zombie invasion 28 days ago!

GENERAL FRED: These zombies... are different. They're... *philosophical* zombies.

CAPTAIN MUDD: Are they filled with rage, causing them to bite people?

COLONEL TODD: Do they lose all capacity for reason?

GENERAL FRED: No. They behave... *exactly* like we do... except that they're not conscious.

(*Silence grips the table.*)

COLONEL TODD: Dear God.

GENERAL FRED moves over to a computerized display.

GENERAL FRED: This is New York City, two weeks ago.

The display shows crowds bustling through the streets, people eating in restaurants, a garbage truck hauling away trash.

GENERAL FRED: *This...* is New York City... *now*.

The display changes, showing a crowded subway train, a group of students laughing in a park, and a couple holding hands in the sunlight.

COLONEL TODD: It's worse than I imagined.

CAPTAIN MUDD: How can you tell, exactly?

COLONEL TODD: I've never seen anything so brutally ordinary.

A lab-coated SCIENTIST stands up at the foot of the table.

SCIENTIST: The zombie disease eliminates consciousness without changing the brain in any way. We've been trying to understand how the disease is transmitted. Our conclusion is that, since the disease attacks dual properties of ordinary matter, it must,

---

<sup>1</sup>. [http://lesswrong.com/lw/pn/zombies\\_the\\_movie/](http://lesswrong.com/lw/pn/zombies_the_movie/)

itself, operate outside our universe. We're dealing with an *epiphenomenal virus*.

GENERAL FRED: Are you sure?

SCIENTIST: As sure as we can be in the total absence of evidence.

GENERAL FRED: All right. Compile a report on every epiphenomenon ever observed. What, where, and who. I want a list of everything that hasn't happened in the last fifty years.

CAPTAIN MUDD: If the virus is epiphenomenal, how do we know it exists?

SCIENTIST: The same way we know *we're* conscious.

CAPTAIN MUDD: Oh, okay.

GENERAL FRED: Have the doctors made any progress on finding an epiphenomenal cure?

SCIENTIST: They've tried every placebo in the book. No dice. Everything they do has an effect.

GENERAL FRED: Have you brought in a homeopath?

SCIENTIST: I tried, sir! I couldn't find any!

GENERAL FRED: Excellent. And the Taoists?

SCIENTIST: They refuse to do anything!

GENERAL FRED: Then we may yet be saved.

COLONEL TODD: What about David Chalmers? Shouldn't he be here?

GENERAL FRED: Chalmers... was one of the first victims.

COLONEL TODD: Oh no.

(Cut to the INTERIOR of a cell, completely walled in by reinforced glass, where DAVID CHALMERS paces back and forth.)

DOCTOR: David! David Chalmers! Can you hear me?

CHALMERS: Yes.

NURSE: It's no use, doctor.

CHALMERS: I'm perfectly fine. I've been introspecting on my consciousness, and I can't detect any difference. I *know* I would be expected to say that, but—

The DOCTOR turns away from the glass screen in horror.

DOCTOR: His words, they... they *don't mean anything*.

CHALMERS: This is a grotesque distortion of my philosophical views. This sort of thing can't actually happen!

DOCTOR: Why not?

NURSE: Yes, why not?

CHALMERS: Because—

(*Cut to two POLICE OFFICERS, guarding a dirt road leading up to the imposing steel gate of a gigantic concrete complex. On their uniforms, a badge reads "BRIDGING LAW ENFORCEMENT AGENCY".*)

OFFICER 1: You've got to watch out for those clever bastards. They look like humans. They can talk like humans. They're identical to humans on the atomic level. But they're not human.

OFFICER 2: Scumbags.

The huge noise of a throbbing engine echoes over the hills. Up rides the MAN on a white motorcycle. The MAN is wearing black sunglasses and a black leather business suit with a black leather tie and silver metal boots. His white beard flows in the wind. He pulls to a halt in front of the gate.

The OFFICERS bustle up to the motorcycle.

OFFICER 1: State your business here.

MAN: Is this where you're keeping David Chalmers?

OFFICER 2: What's it to you? You a friend of his?

MAN: Can't say I am. But even zombies have rights.

OFFICER 1: All right, buddy, let's see your qualia.

MAN: I don't have any.

OFFICER 2 suddenly pulls a gun, keeping it trained on the MAN. OFFICER 2: Aha! A zombie!

OFFICER 1: No, zombies claim to have qualia.

OFFICER 2: So he's an ordinary human?

OFFICER 1: No, they also claim to have qualia.

The OFFICERS look at the MAN, who waits calmly.

OFFICER 2: Um...

OFFICER 1: Who *are* you?

MAN: I'm Daniel Dennett, bitches.

Seemingly from nowhere, DENNETT pulls a sword and slices OFFICER 2's gun in half with a steely noise. OFFICER 1 begins to

reach for his own gun, but DENNETT is suddenly standing behind OFFICER 1 and chops with a fist, striking the junction of OFFICER 1's shoulder and neck. OFFICER 1 drops to the ground.

OFFICER 2 steps back, horrified.

OFFICER 2: That's not possible! How'd you do that?

DENNETT: I am one with my body.

DENNETT drops OFFICER 2 with another blow, and strides toward the gate. He looks up at the imposing concrete complex, and grips his sword tighter.

DENNETT (*quietly to himself*): There is a spoon.

(Cut back to GENERAL FRED and the other military officials.)

GENERAL FRED: I've just received the reports. We've lost Detroit.

CAPTAIN MUDD: I don't want to be the one to say "Good riddance", but—

GENERAL FRED: Australia has been... *reduced to atoms*.

COLONEL TODD: The epiphenomenal virus is spreading faster. Civilization itself threatens to dissolve into total normality. We could be looking at the middle of humanity.

CAPTAIN MUDD: Can we negotiate with the zombies?

GENERAL FRED: We've sent them messages. They sent only a single reply.

CAPTAIN MUDD: Which was...?

GENERAL FRED: It's on its way now.

An orderly brings in an envelope, and hands it to GENERAL FRED.

GENERAL FRED opens the envelope, takes out a single sheet of paper, and reads it.

Silence envelops the room.

CAPTAIN MUDD: What's it say?

GENERAL FRED: It says... that *we're* the ones with the virus.  
(A silence falls.)

COLONEL TODD raises his hands and stares at them.

COLONEL TODD: My God, it's true. It's true. I...

(A tear rolls down COLONEL TODD's cheek.)

COLONEL TODD: I don't feel anything.

The screen goes black.  
The sound goes silent.  
The movie continues exactly as before.

---

Elizombies

PS: This is  
me being at-  
tacked by  
zombie nurses  
at Penguincon.

Only at a  
*combination* sci-  
ence fiction  
and open-  
source  
convention  
would it be  
possible to at-  
tend a session

on knife-throwing, cry “In the name of Bayes, die!”, throw the knife,  
and then have a fellow holding a wooden shield say, “Yes, but how  
do you determine the prior for where the knife hits?”

## 35. Excluding the Supernatural<sup>1</sup>

**Followup to:** Reductionism<sup>2</sup>, Anthropomorphic Optimism<sup>3</sup>

Occasionally, you hear someone claiming that creationism should not be taught in schools, especially not as a competing hypothesis to evolution, because creationism is *a priori* and automatically excluded from scientific consideration, in that it invokes the “supernatural”.

So... is the idea here, that creationism *could* be true, but *even if it were true*, you wouldn’t be *allowed* to teach it in science class, because science is only about “natural” things?

It seems clear enough that this notion stems from the desire to avoid a confrontation between science and religion<sup>4</sup>. You don’t want to come right out and say that science doesn’t teach Religious Claim X because X has been tested by the scientific method and found false<sup>5</sup>. So instead, you can... um... claim that science is excluding hypothesis X *a priori*. That way you don’t have to discuss how experiment has falsified X *a posteriori*.

Of course this plays right into the creationist claim that Intelligent Design isn’t getting a fair shake from science—that science has *prejudged* the issue in favor of atheism, regardless of the evidence. If science excluded Intelligent Design *a priori*, this would be a justified complaint!

But let’s back up a moment. The one comes to you and says: “Intelligent Design is excluded from being science *a priori*, because it is ‘supernatural’, and science only deals in ‘natural’ explanations.”

What exactly do they mean, “supernatural”? Is any explanation invented by someone with the last name “Cohen” a supernatural one? If we’re going to summarily kick a set of hypotheses out of science, what is it that we’re supposed to exclude?

By *far* the best definition I’ve ever heard of the supernatural is Richard Carrier’s<sup>6</sup>: A “supernatural” explanation appeals to *ontolog-*

---

1. [http://lesswrong.com/lw/tv/excluding\\_the\\_supernatural/](http://lesswrong.com/lw/tv/excluding_the_supernatural/)

2. Page 699, ‘Reductionism’.

3. [http://lesswrong.com/lw/st/anthropomorphic\\_optimism/](http://lesswrong.com/lw/st/anthropomorphic_optimism/)

4. [http://lesswrong.com/lw/i8/religious\\_claim\\_to\\_be\\_nondisprovable/](http://lesswrong.com/lw/i8/religious_claim_to_be_nondisprovable/)

5. [http://lesswrong.com/lw/i8/religious\\_claim\\_to\\_be\\_nondisprovable/](http://lesswrong.com/lw/i8/religious_claim_to_be_nondisprovable/)

6. <http://richardcarrier.blogspot.com/2007/01/defining-supernatural.html>

*ically basic mental things*, mental entities that cannot be reduced to nonmental entities.

This is the difference, for example, between saying that water rolls downhill because it *wants* to be lower<sup>7</sup>, and setting forth differential equations that claim to describe only motions, not desires. It's the difference between saying that a tree puts forth leaves because of a tree spirit, versus examining plant biochemistry. Cognitive science takes the fight against supernaturalism into the realm of the mind<sup>8</sup>.

Why is this an excellent definition of the supernatural? I refer you to Richard Carrier<sup>9</sup> for the full argument. But consider: Suppose that you discover what seems to be a *spirit*, inhabiting a tree: a dryad who can materialize outside or inside the tree, who speaks in English about the need to protect her tree, et cetera. And then suppose that we turn a microscope on this tree spirit, and she turns out to be made of parts<sup>10</sup>—not inherently spiritual and ineffable parts, like fabric of desireness and cloth of belief; but rather the same sort of parts as quarks and electrons, parts whose behavior is defined in motions rather than minds. Wouldn't the dryad immediately be demoted to the dull catalogue of common things<sup>11</sup>?

But if we accept Richard Carrier's definition of the supernatural, then a dilemma arises: we *want* to give religious claims a fair shake, but it seems that we have *very good* grounds for excluding supernatural explanations *a priori*.

I mean, what *would* the universe look like if reductionism were false?

I previously defined the reductionist thesis<sup>12</sup> as follows: human minds create multi-level *models* of reality in which high-level patterns and low-level patterns are separately and explicitly *represented*. A physicist knows Newton's equation for gravity, Einstein's equation for gravity, and the derivation of the former as a low-speed approximation of the latter. But these three separate mental representations, are only a convenience of human cognition. It is

7. [http://lesswrong.com/lw/te/three\\_fallacies\\_of\\_teachology/](http://lesswrong.com/lw/te/three_fallacies_of_teachology/)

8. [http://lesswrong.com/lw/tf/dreams\\_of\\_ai\\_design/](http://lesswrong.com/lw/tf/dreams_of_ai_design/)

9. <http://richardcarrier.blogspot.com/2007/01/defining-supernatural.html>

10. [http://lesswrong.com/lw/t5/when\\_antropomorphism\\_became\\_stupid/](http://lesswrong.com/lw/t5/when_antropomorphism_became_stupid/)

11. Page 715, 'Joy in the Merely Real'.

12. Page 699, 'Reductionism'.

not that *reality itself* has an Einstein equation that governs at high speeds, a Newton equation that governs at low speeds, and a “bridging law” that smooths the interface. Reality itself has only a single level, Einsteinian gravity. It is only the Mind Projection Fallacy<sup>13</sup> that makes some people talk as if the higher levels could have a separate existence—different levels of organization can have separate representations in human maps, but the territory itself is a single unified low-level mathematical object.

Suppose this were wrong.

Suppose that the Mind Projection Fallacy<sup>14</sup> was not a fallacy, but simply true.

Suppose that a 747 had a fundamental physical existence apart from the quarks making up the 747.

What experimental observations would you expect to make, if you found yourself in such a universe?

If you can’t come up with a good answer to that, it’s not *observation* that’s ruling out “non-reductionist” beliefs, but *a priori* logical incoherence. If you can’t say what predictions the “non-reductionist” model makes, how can you say that experimental evidence rules it out?

My thesis is that non-reductionism is a *confusion*; and once you realize that an idea is a confusion, it becomes a tad difficult to envision what the universe would look like if the confusion were *true*. Maybe I’ve got some multi-level model of the world, and the multi-level model has a one-to-one direct correspondence with the causal elements of the physics? But once all the rules are specified, why wouldn’t the model just flatten out into yet another list of fundamental things and their interactions? Does everything I can *see in* the model, like a 747 or a human mind, have to become a separate real thing? But what if I see a pattern in that new supersystem?

Supernaturalism is a special case of non-reductionism, where it is not 747s that are irreducible, but just (some) mental things. Religion is a special case of supernaturalism, where the irreducible mental things are God(s) and souls; and perhaps also sins, angels, karma, etc.

---

<sup>13</sup>. Page 685, ‘Mind Projection Fallacy’.

<sup>14</sup>. Page 685, ‘Mind Projection Fallacy’.

If I propose the existence of a powerful entity with the ability to survey and alter each element of our observed universe, but with the entity reducible to nonmental parts that interact with the elements of our universe in a lawful way; if I propose that this entity wants certain particular things, but “wants” using a brain composed of particles and fields; then this is not yet a religion, just a naturalistic hypothesis about a naturalistic Matrix. If tomorrow the clouds parted and a vast glowing amorphous figure thundered forth the above description of reality, then this would not imply that the figure was necessarily honest; but I would show the movies in a science class, and I would try to derive testable predictions from the theory.

Conversely, religions have ignored<sup>15</sup> the discovery of that ancient bodiless thing<sup>16</sup>: omnipresent in the working of Nature and immanent in every falling leaf: vast as a planet’s surface and billions of years old: itself unmade and arising from the structure of physics: designing without brain to shape all life on Earth and the minds of humanity. Natural selection, when Darwin proposed it, was not hailed as the long-awaited Creator: It wasn’t *fundamentally* mental.

But now we get to the dilemma: if the staid conventional normal boring understanding of physics and the brain *is* correct, there’s no way *in principle* that a human being can concretely envision, and derive testable experimental predictions about, an alternate universe in which things *are* irreducibly mental. Because, if the boring old normal model is correct, your brain is made of quarks, and so your brain will only be able to envision and concretely predict things that can predicted by quarks. You will only ever be able to construct models made of interacting simple things.

People who live in reductionist universes cannot concretely envision non-reductionist universes. They can pronounce the syllables<sup>17</sup> “non-reductionist” but they can’t *imagine* it.

The basic error of anthropomorphism, and the reason why supernatural explanations sound much simpler than they really are<sup>18</sup>, is your brain using itself as an opaque black box to predict other things labeled “mindful”. Because you already have big, complicat-

---

<sup>15</sup>. [http://lesswrong.com/lw/kr/an\\_alien\\_god/](http://lesswrong.com/lw/kr/an_alien_god/)

<sup>16</sup>. [http://lesswrong.com/lw/kr/an\\_alien\\_god/](http://lesswrong.com/lw/kr/an_alien_god/)

<sup>17</sup>. Page 114, ‘Professing and Cheering’.

<sup>18</sup>. Page 95, ‘Occam’s Razor’.

ed webs of neural circuitry that implement your “wanting” things, it seems like you can easily describe water that “wants” to flow downhill—the one word “want” acts as a lever<sup>19</sup> to set your *own* complicated wanting-machinery in motion.

Or you imagine that God likes beautiful things, and therefore made the flowers. Your own “beauty” circuitry determines what is “beautiful” and “not beautiful”. But you don’t know the diagram of your own synapses. You can’t describe a *nonmental* system that computes the same label for what is “beautiful” or “not beautiful”—can’t write a computer program that predicts your own labelings. But this is just a defect of knowledge on your part; it doesn’t mean that the brain has no explanation<sup>20</sup>.

If the “boring view” of reality is correct, then you can *never* predict anything irreducible because *you* are reducible. You can never get Bayesian confirmation for a hypothesis of irreducibility, because any *prediction you can make* is, therefore, something that could also be predicted by a reducible thing, namely your brain.

Some boxes you really *can’t* think outside. If our universe *really is* Turing computable, we will never be able to *concretely* envision anything that isn’t Turing-computable—no matter how many levels of halting oracle hierarchy our mathematicians can talk *about*, we won’t be able to predict what a halting oracle would actually *say*, in such fashion as to experimentally discriminate it from merely computable reasoning.

Of course, that’s all assuming the “boring view” is correct. *To the extent* that you believe evolution is true, you should not expect to encounter strong evidence against evolution. To the extent you believe reductionism is true, you should expect non-reductionist hypotheses to be *incoherent* as well as wrong. To the extent you believe supernaturalism is false, you should expect it to be *inconceivable* as well.

If, on the other hand, a supernatural hypothesis turns out to be true, then presumably you will also discover that it is not inconceivable.

So let us bring this back full circle to the matter of Intelligent Design:

---

19. [http://lesswrong.com/lw/sp/detached\\_lever\\_fallacy/](http://lesswrong.com/lw/sp/detached_lever_fallacy/)

20. Page 154, ‘Mysterious Answers to Mysterious Questions’.

Should ID be excluded *a priori* from experimental falsification and science classrooms, because, by invoking the supernatural, it has placed itself outside of natural philosophy?

I answer: “Of course not.” The *irreducibility* of the intelligent designer is not an indispensable part of the ID hypothesis. For every irreducible God that can be proposed by the IDers, there exists a corresponding reducible alien that behaves in accordance with the same predictions—since the IDers themselves are reducible; to the extent I believe reductionism is in fact correct, which is a rather strong extent, I must expect to discover reducible formulations of all supposedly supernatural predictive models.

If we’re going over the archeological records to test the assertion that Jehovah parted the Red Sea out of an explicit desire to display its superhuman power, then it makes little difference whether Jehovah is ontologically basic, or an alien with nanotech, or a Dark Lord of the Matrix. You do some archeology, find no skeletal remnants or armor at the Red Sea site, and indeed find records that Egypt ruled much of Canaan at the time. So you stamp the historical record in the Bible “disproven” and carry on. The hypothesis is coherent, falsifiable and wrong.

Likewise with the evidence from biology that foxes are designed to chase rabbits, rabbits are designed to evade foxes, and neither is designed “to carry on their species” or “protect the harmony of Nature”<sup>21</sup>; likewise with the retina being designed backwards with the light-sensitive parts at the bottom; and so on through a thousand other items of evidence for splintered, immoral, incompetent<sup>22</sup> design. The Jehovah model of our alien god<sup>23</sup> is coherent, falsifiable, and wrong—coherent, that is, so long as you don’t care whether Jehovah is ontologically basic or just an alien.

Just convert the supernatural hypothesis into the corresponding natural hypothesis. Just make the same predictions the same way, without asserting any mental things to be ontologically basic. Consult your brain’s black box if necessary to make predictions—say, if you want to talk about an “angry god” without building a full-fledged angry AI to label behaviors as angry or not angry. So you derive the predictions, or look up the predictions made by ancient

21. [http://lesswrong.com/lw/l5/evolving\\_to\\_extinction/](http://lesswrong.com/lw/l5/evolving_to_extinction/)

22. [http://lesswrong.com/lw/kt/evolutions\\_are\\_stupid\\_but\\_work\\_anyway/](http://lesswrong.com/lw/kt/evolutions_are_stupid_but_work_anyway/)

23. [http://lesswrong.com/lw/kr/an\\_alien\\_god/](http://lesswrong.com/lw/kr/an_alien_god/)

theologians without advance knowledge of our experimental results. If experiment conflicts with those predictions, then it is fair to speak of the religious claim having been scientifically refuted. It was given its just chance at confirmation; it is being excluded *a posteriori*, not *a priori*.

Ultimately, reductionism is just disbelief in *fundamentally complicated* things. If “fundamentally complicated” sounds like an oxymoron... well, that’s why I think that the doctrine of non-reductionism is a *confusion*, rather than a way that things could be, but aren’t. You would be wise to be wary, if you find yourself supposing such things.

But the ultimate rule of science is to look and see. If ever a God appeared to thunder upon the mountains, it would be something that people looked at and saw.

*Corollary:* Any supposed designer<sup>24</sup> of Artificial General Intelligence who talks about religious beliefs in respectful tones<sup>25</sup>, is clearly not an expert on<sup>26</sup> reducing mental things to nonmental things; and indeed knows so very little of the uttermost basics, as for it to be scarcely plausible that they could be expert at<sup>27</sup> the art; unless their *idiot savancy* is complete. Or, of course, if they’re outright lying. We’re not talking about a subtle mistake.

---

24. [http://lesswrong.com/lw/tf/dreams\\_of\\_ai\\_design/](http://lesswrong.com/lw/tf/dreams_of_ai_design/)

25. [http://lesswrong.com/lw/gv/outside\\_the\\_laboratory/](http://lesswrong.com/lw/gv/outside_the_laboratory/)

26. [http://www.overcomingbias.com/2007/04/expert\\_at\\_versu.html](http://www.overcomingbias.com/2007/04/expert_at_versu.html)

27. [http://www.overcomingbias.com/2007/04/expert\\_at\\_versu.html](http://www.overcomingbias.com/2007/04/expert_at_versu.html)

## 36. Psychic Powers<sup>1</sup>

**Followup to:** Excluding the Supernatural<sup>2</sup>

Yesterday, I wrote:

If the “boring view” of reality is correct, then you can *never* predict anything irreducible because *you* are reducible. You can never get Bayesian confirmation for a hypothesis of irreducibility, because any *prediction you can make* is, therefore, something that could also be predicted by a reducible thing, namely your brain.

Benja Fallenstein commented<sup>3</sup>:

I think that while you can in this case never devise an empirical test whose outcome could *logically prove* irreducibility, there is no clear reason to believe that you cannot devise a test whose counterfactual outcome in an irreducible world would make irreducibility subjectively *much more probable* (given an Occamian prior).

Without getting into reducibility/irreducibility, consider the scenario that the physical universe makes it possible to build a hypercomputer—that performs operations on arbitrary real numbers, for example—but that our brains do not actually make use of this: they can be simulated perfectly well by an ordinary Turing machine, thank you very much...

Well, that’s a very intelligent argument, Benja Fallenstein. But I have a crushing reply to your argument, such that, once I deliver it, you will at once give up further debate with me on this particular point:

You’re right.

---

1. [http://lesswrong.com/lw/tw/psychic\\_powers/](http://lesswrong.com/lw/tw/psychic_powers/)

2. Page 826, ‘Excluding the Supernatural’.

3. <http://www.overcomingbias.com/2008/09/excluding-the-s.html#comment-130388082>

Alas, I don't get modesty credit on this one, because after publishing yesterday's post I realized a similar flaw on my own—this one concerning Occam's Razor and psychic powers:

If beliefs and desires are irreducible and ontologically basic entities, or have an ontologically basic *component* not covered by existing science, that would make it far more likely that there was an ontological rule governing the interaction of different minds—an interaction which bypassed ordinary “material” means of communication like sound waves, known to existing science.

If naturalism is correct, then there exists a conjugate reductionist model that makes the *same predictions* as any concrete prediction that any parapsychologist can make about telepathy.

Indeed, if naturalism is correct, the only reason we can *conceive* of beliefs as “fundamental” is due to lack of self-knowledge of our own neurons—that the peculiar reflective architecture of our own minds exposes the “belief” class<sup>4</sup> but hides the machinery behind it<sup>5</sup>.

Nonetheless, the discovery of information transfer between brains, in the absence of any known material connection between them, is *probabilistically* a privileged prediction of supernatural models (those that contain ontologically basic mental entities). Just because it is so much *simpler* in that case to have a new law relating beliefs between different minds, compared to the “boring” model where beliefs are complex constructs of neurons.

The hope of psychic powers arises from treating beliefs and desires as sufficiently fundamental objects that they can have *unmediated* connections to reality. If beliefs are patterns of neurons made of known material, with inputs given by organs like eyes constructed of known material, and with outputs through muscles constructed of known material, and this seems sufficient to account for all known mental powers of humans, then there's no reason to expect anything more—no reason to postulate additional connections. This is why reductionists don't expect psychic powers. Thus, observing psychic powers would be strong evidence for the supernatural in Richard Carrier's<sup>6</sup> sense.

---

4. [http://lesswrong.com/lw/sp/detached\\_lever\\_fallacy/](http://lesswrong.com/lw/sp/detached_lever_fallacy/)

5. [http://lesswrong.com/lw/sp/detached\\_lever\\_fallacy/](http://lesswrong.com/lw/sp/detached_lever_fallacy/)

6. Page 826, ‘Excluding the Supernatural’.

We have an Occam rule that counts the number of ontologically basic classes and ontologically basic laws in the model, and penalizes the count of entities. If naturalism is correct, then the attempt to count “belief” or the “relation between belief and reality” as a single basic entity, is simply misguided anthropomorphism; we are only tempted to it by a quirk of our brain’s internal architecture. But if you *just go with* that misguided view, then it assigns a much higher probability to psychic powers than does naturalism, because you can implement psychic powers using apparently simpler laws.

Hence the actual discovery of psychic powers would imply that the human-naive Occam rule was in fact better-calibrated than the sophisticated naturalistic Occam rule. It would argue that reductionists had been wrong all along in trying to take apart the brain; that what our minds exposed as a seemingly simple lever, was in fact a simple lever. The naive dualists would have been right from the beginning, which is why their ancient wish would have been enabled to come true.

So telepathy, and the ability to influence events just by wishing at them, and precognition, would all, if discovered, be strong Bayesian evidence in favor of the hypothesis that beliefs are ontologically fundamental. Not logical proof, but strong Bayesian evidence.

If reductionism is correct, then any science-fiction story containing psychic powers, can be output by a system of simple elements (i.e., the story’s author’s brain); but if we *in fact* discover psychic powers, that would make it much more probable that events were occurring which could not *in fact* be described by reductionist models.

Which just goes to say: The existence of psychic powers is a privileged probabilistic assertion of non-reductionist worldviews—they *own* that advance prediction; they devised it and put it forth, in defiance of reductionist expectations. So by the laws of science, if psychic powers are discovered, non-reductionism wins.

I am therefore confident<sup>7</sup> in dismissing psychic powers as *a priori* implausible, despite all the claimed experimental evidence in favor of them.

---

7. [http://lesswrong.com/lw/ig/i\\_defy\\_the\\_data/](http://lesswrong.com/lw/ig/i_defy_the_data/)



## **Part VI**

### **The Metaethics Sequence**

*What words like “right” and “should” mean; how to integrate moral concepts into a naturalistic universe.*

*(The dependencies on this sequence may not be fully organized, and the post list does not have summaries. Yudkowsky considers this one of his less successful attempts at explanation.)*



# I. Heading Toward Morality<sup>1</sup>

**Followup to:** Ghosts in the Machine<sup>2</sup>, Fake Fake Utility Functions<sup>3</sup>, Fake Utility Functions<sup>4</sup>

As people were complaining before about not seeing where the quantum physics sequence<sup>5</sup> was going, I shall go ahead and tell you where I'm heading now.

Having dissolved<sup>6</sup> the confusion surrounding the word “could”<sup>7</sup>, the trajectory is now heading toward *should*.

In fact, I've been heading there for a while. Remember the whole sequence<sup>8</sup> on fake utility functions<sup>9</sup>? Back in... well... November 2007?

I sometimes think of there being a train that goes to the Friendly AI station; but it makes several stops before it gets there; and at each stop, a large fraction of the remaining passengers get off.

One of those stops is the one I spent a month leading up to in November 2007, the sequence chronicled in Fake Fake Utility Functions<sup>10</sup> and concluded in Fake Utility Functions<sup>11</sup>.

That's the stop where someone thinks of the One Great Moral Principle That Is All We Need To Give AIs.

To deliver that one warning, I had to go through all sorts of topics—which topics one might find useful even if not working on Friendly AI. I warned against Affective<sup>12</sup> Death<sup>13</sup> Spirals<sup>14</sup>, which required recursing on the affect heuristic<sup>15</sup> and halo<sup>16</sup> ef-

---

1. [http://lesswrong.com/lw/rb/heading\\_toward\\_morality/](http://lesswrong.com/lw/rb/heading_toward_morality/)

2. [http://lesswrong.com/lw/rf/ghosts\\_in\\_the\\_machine/](http://lesswrong.com/lw/rf/ghosts_in_the_machine/)

3. [http://lesswrong.com/lw/lp/fake\\_fake\\_utility\\_functions/](http://lesswrong.com/lw/lp/fake_fake_utility_functions/)

4. [http://lesswrong.com/lw/lq/fake\\_utility\\_functions/](http://lesswrong.com/lw/lq/fake_utility_functions/)

5. [http://lesswrong.com/lw/r5/the\\_quantum\\_physics\\_sequence/](http://lesswrong.com/lw/r5/the_quantum_physics_sequence/)

6. Page 673, ‘Dissolving the Question’.

7. [http://lesswrong.com/lw/rb/possibility\\_and\\_couldness/](http://lesswrong.com/lw/rb/possibility_and_couldness/)

8. [http://lesswrong.com/lw/lp/fake\\_fake\\_utility\\_functions/](http://lesswrong.com/lw/lp/fake_fake_utility_functions/)

9. [http://lesswrong.com/lw/lq/fake\\_utility\\_functions/](http://lesswrong.com/lw/lq/fake_utility_functions/)

10. [http://lesswrong.com/lw/lp/fake\\_fake\\_utility\\_functions/](http://lesswrong.com/lw/lp/fake_fake_utility_functions/)

11. [http://lesswrong.com/lw/lq/fake\\_utility\\_functions/](http://lesswrong.com/lw/lq/fake_utility_functions/)

12. Page 384, ‘Affective Death Spirals’.

13. Page 386, ‘Resist the Happy Death Spiral’.

14. Page 392, ‘Uncritical Supercriticality’.

15. Page 363, ‘The Affect Heuristic’.

fect<sup>17</sup>, so that your good feeling about one particular moral principle wouldn't spiral out of control. I did that<sup>18</sup> whole<sup>19</sup> sequence<sup>20</sup> on evolution<sup>21</sup>; and discursed on the human ability to make almost any goal appear to support almost any policy<sup>22</sup>; I went into evolutionary psychology<sup>23</sup> to argue for why we shouldn't expect human terminal values<sup>24</sup> to reduce to any simple principle<sup>25</sup>, even happiness<sup>26</sup>, explaining the concept of "expected utility"<sup>27</sup> along the way...

...and talked about genies<sup>28</sup> and more; but you can read the Fake Utility sequence<sup>29</sup> for that.

So that's *just* the warning against trying to oversimplify human morality<sup>30</sup> into One Great Moral Principle.

If you want to actually dissolve the confusion<sup>31</sup> that surrounds the word "should"—which is the next stop on the train—then that takes a much longer introduction. Not just one November.

I went through the sequence on words and definitions<sup>32</sup> so that I would be able to later say things like "The next project is to Taboo<sup>33</sup> the word 'should' and replace it with its substance<sup>34</sup>", or "Sorry, saying that morality is self-interest 'by definition'<sup>35</sup> isn't going to cut it here".

And also the words-and-definitions sequence was the simplest example I knew to introduce the notion of How An Algorithm

16. Page 374, 'The Halo Effect'.

17. Page 378, 'Superhero Bias'.

18. [http://lesswrong.com/lw/kr/an\\_alien\\_god/](http://lesswrong.com/lw/kr/an_alien_god/)

19. [http://lesswrong.com/lw/ks/the\\_wonder\\_of\\_evolution/](http://lesswrong.com/lw/ks/the_wonder_of_evolution/)

20. [http://lesswrong.com/lw/kt/evolutions\\_are\\_stupid\\_but\\_work\\_anyway/](http://lesswrong.com/lw/kt/evolutions_are_stupid_but_work_anyway/)

21. [http://lesswrong.com/lw/kw/the\\_tragedy\\_of\\_group\\_selectionism/](http://lesswrong.com/lw/kw/the_tragedy_of_group_selectionism/)

22. Page 538, 'Fake Optimization Criteria'.

23. [http://lesswrong.com/lw/l1/evolutionary\\_psychology/](http://lesswrong.com/lw/l1/evolutionary_psychology/)

24. [http://lesswrong.com/lw/l4/terminal\\_values\\_and\\_instrumental\\_values/](http://lesswrong.com/lw/l4/terminal_values_and_instrumental_values/)

25. [http://lesswrong.com/lw/l3/thou\\_art\\_godshatter/](http://lesswrong.com/lw/l3/thou_art_godshatter/)

26. [http://lesswrong.com/lw/lb/not\\_for\\_the\\_sake\\_of\\_happiness\\_alone/](http://lesswrong.com/lw/lb/not_for_the_sake_of_happiness_alone/)

27. [http://lesswrong.com/lw/l4/terminal\\_values\\_and\\_instrumental\\_values/](http://lesswrong.com/lw/l4/terminal_values_and_instrumental_values/)

28. [http://lesswrong.com/lw/ld/the\\_hidden\\_complexity\\_of\\_wishes/](http://lesswrong.com/lw/ld/the_hidden_complexity_of_wishes/)

29. [http://lesswrong.com/lw/lp/fake\\_fake\\_utility\\_functions/](http://lesswrong.com/lw/lp/fake_fake_utility_functions/)

30. [http://lesswrong.com/lw/lq/fake\\_utility\\_functions/](http://lesswrong.com/lw/lq/fake_utility_functions/)

31. Page 673, 'Dissolving the Question'.

32. Page 301, '37 Ways That Words Can Be Wrong'.

33. Page 245, 'Taboo Your Words'.

34. Page 248, 'Replace the Symbol with the Substance'.

35. Page 263, 'Arguing "By Definition"'.

Feels From Inside<sup>36</sup>, which is one of the great master keys to dissolving<sup>37</sup> wrong questions<sup>38</sup>. Though it seems to us that our cognitive representations are the very substance of the world<sup>39</sup>, they have a character that comes from cognition and often cuts cross-wise to a universe made of quarks. E.g. probability<sup>40</sup>; if we are uncertain of a phenomenon, that is a fact about our state of mind, not an intrinsic character of the phenomenon.

Then the reductionism sequence: that a universe made only of quarks<sup>41</sup>, does not mean that things of value are lost<sup>42</sup> or even degraded to mundanity<sup>43</sup>. And the notion of how the sum can seem unlike the parts<sup>44</sup>, and yet be as much the parts as our hands are fingers<sup>45</sup>.

Followed by a new example, one step up in difficulty from words and their seemingly intrinsic meanings<sup>46</sup>: “Free will” and seemingly intrinsic could-ness<sup>47</sup>.

But before that point, it was useful to introduce quantum physics<sup>48</sup>. Not just to get to timeless physics<sup>49</sup> and dissolve the “determinism<sup>50</sup>” part of the “free will” confusion. But also, more fundamentally, to break belief in an intuitive universe<sup>51</sup> that looks just like our brain’s cognitive representations. And present examples of the dissolution of even such fundamental intuitions as those concerning personal identity<sup>52</sup>. And to illustrate the idea that you

---

36. Page 223, ‘How An Algorithm Feels From Inside’.

37. Page 673, ‘Dissolving the Question’.

38. Page 679, ‘Wrong Questions’.

39. Page 685, ‘Mind Projection Fallacy’.

40. Page 687, ‘Probability is in the Mind’.

41. Page 699, ‘Reductionism’.

42. Page 703, ‘Explaining vs. Explaining Away’.

43. Page 715, ‘Joy in the Merely Real’.

44. Page 759, ‘Heat vs. Motion’.

45. Page 751, ‘Hand vs. Fingers’.

46. Page 234, ‘Feel the Meaning’.

47. [http://lesswrong.com/lw/rb/possibility\\_and\\_couldness/](http://lesswrong.com/lw/rb/possibility_and_couldness/)

48. [http://lesswrong.com/lw/r5/the\\_quantum\\_physics\\_sequence/](http://lesswrong.com/lw/r5/the_quantum_physics_sequence/)

49. [http://lesswrong.com/lw/qp/timeless\\_physics/](http://lesswrong.com/lw/qp/timeless_physics/)

50. [http://lesswrong.com/lw/r1/timeless\\_control/](http://lesswrong.com/lw/r1/timeless_control/)

51. [http://lesswrong.com/lw/qy/why\\_quantum/](http://lesswrong.com/lw/qy/why_quantum/)

52. [http://lesswrong.com/lw/qx/timeless\\_identity/](http://lesswrong.com/lw/qx/timeless_identity/)

are within physics<sup>53</sup>, within causality<sup>54</sup>, and that strange things will go wrong in your mind if ever you forget it.

Lately we have begun to approach the final precautions, with warnings against such notions as Author\* control<sup>55</sup>: every mind which computes a morality must do so within a chain of lawful causality, it cannot arise from the free will of a ghost in the machine<sup>56</sup>.

And the warning against Passing the Recursive Buck<sup>57</sup> to some meta-morality that is not itself computably specified, or some meta-morality that is chosen by a ghost without it being programmed in, or to a notion of “moral truth” just as confusing as “should” itself...

And the warning on the difficulty of grasping slippery things<sup>58</sup> like “should”—demonstrating how very easy it will be to just invent another black box equivalent to should-ness, to sweep should-ness under a slightly different rug—or to bounce off into mere modal logics of primitive should-ness...

We aren’t yet at the point where I can explain morality.

But I think—though I could be mistaken—that we are finally getting close to the final sequence.

And if you don’t care about my goal of explanatorily transforming Friendly AI from a Confusing Problem into a merely Extremely Difficult Problem, then stick around anyway. I tend to go through interesting intermediates along my way.

It might seem like confronting “the nature of morality” from the perspective of Friendly AI is only asking for additional trouble.

Artificial Intelligence melts people’s brains. Metamorality melts people’s brains. Trying to think about AI and metamorality at the same time can cause people’s brains to spontaneously combust and burn for years, emitting toxic smoke—don’t laugh, I’ve seen it happen multiple times.

But the discipline imposed by Artificial Intelligence is this: you cannot escape into things that are “self-evident” or “obvious”. That

53. [http://lesswrong.com/lw/ro/thou\\_art\\_physics/](http://lesswrong.com/lw/ro/thou_art_physics/)

54. [http://lesswrong.com/lw/ra/causality\\_and\\_moral\\_responsibility/](http://lesswrong.com/lw/ra/causality_and_moral_responsibility/)

55. [http://lesswrong.com/lw/rc/the\\_ultimate\\_source/](http://lesswrong.com/lw/rc/the_ultimate_source/)

56. [http://lesswrong.com/lw/rf/ghosts\\_in\\_the\\_machine/](http://lesswrong.com/lw/rf/ghosts_in_the_machine/)

57. [http://lesswrong.com/lw/rd/passing\\_the\\_recursive\\_buck/](http://lesswrong.com/lw/rd/passing_the_recursive_buck/)

58. [http://lesswrong.com/lw/re/grasping\\_slippery\\_things/](http://lesswrong.com/lw/re/grasping_slippery_things/)

doesn't stop people from trying, but the programs don't work. Every thought has to be computed somehow, by transistors made of mere quarks, and not by moral self-evidence to some ghost in the machine.

If what you care about is rescuing children from burning orphanages, I don't think you will find many moral surprises here; my metamorality adds up to moral normality, as it should<sup>59</sup>. You do not need to worry about metamorality when you are *personally* trying to rescue children from a burning orphanage. The point at which metamoral issues *per se* have high stakes in the real world, is when you try to compute morality in an AI standing in front of a burning orphanage.

Yet there is also a good deal of needless despair and misguided fear of science, stemming from notions such as, "Science tells us the universe is empty of morality". This is damage done by a confused metamorality that fails to add up to moral normality. For that I hope to write down a counterspell of understanding. Existential depression has always annoyed me; it is one of the world's most pointless forms of suffering.

Don't expect the final post on this topic to come tomorrow, but at least you know where we're heading.

---

59. <http://www.singinst.org/blog/2007/06/16/transhumanism-as-simplified-humanism/>

## 2. No Universally Compelling Arguments<sup>1</sup>

**Followup to:** The Design Space of Minds-in-General<sup>2</sup>, Ghosts in the Machine<sup>3</sup>, A Priori<sup>4</sup>

What is so *terrifying* about the idea that not every possible mind might agree with us, even in principle?

For some folks, nothing—it doesn't bother them in the slightest. And for some of *those* folks, the *reason* it doesn't bother them is that they don't have strong intuitions about standards and truths that go beyond personal whims. If they say the sky is blue, or that murder is wrong, that's just their personal opinion; and that someone else might have a different opinion doesn't surprise them.

For other folks, a disagreement that persists even *in principle* is something they can't accept. And for some of *those* folks, the *reason* it bothers them, is that it seems to them that if you allow that some people cannot be persuaded *even in principle* that the sky is blue, then you're conceding that “the sky is blue” is merely an *arbitrary* personal opinion.

Yesterday<sup>5</sup>, I proposed that you should resist the temptation to generalize over all of mind design space. If we restrict ourselves to minds specifiable in a trillion bits or less, then each *universal* generalization “All minds m: X(m)” has two to the trillionth chances to be false, while each *existential* generalization “Exists mind m: X(m)” has two to the trillionth chances to be true.

This would seem to argue that for every argument A, howsoever convincing it may seem to us, there exists at least one possible mind that doesn't buy it.

And the surprise and/or horror of this prospect (for some) has a great deal to do, I think, with the intuition of the ghost-in-the-machine<sup>6</sup>—a ghost with some irreducible core that any *truly valid* argument will convince.

---

1. [http://lesswrong.com/lw/rn/no\\_universally\\_compelling\\_arguments/](http://lesswrong.com/lw/rn/no_universally_compelling_arguments/)

2. [http://lesswrong.com/lw/rm/the\\_design\\_space\\_of\\_minds\\_in\\_general/](http://lesswrong.com/lw/rm/the_design_space_of_minds_in_general/)

3. [http://lesswrong.com/lw/rf/ghosts\\_in\\_the\\_machine/](http://lesswrong.com/lw/rf/ghosts_in_the_machine/)

4. [http://lesswrong.com/lw/kz/a\\_priori/](http://lesswrong.com/lw/kz/a_priori/)

5. [http://lesswrong.com/lw/rm/the\\_design\\_space\\_of\\_minds\\_in\\_general/](http://lesswrong.com/lw/rm/the_design_space_of_minds_in_general/)

6. [http://lesswrong.com/lw/rf/ghosts\\_in\\_the\\_machine/](http://lesswrong.com/lw/rf/ghosts_in_the_machine/)

I have previously spoken<sup>7</sup> of the intuition whereby people map *programming a computer*, onto *instructing a human servant*, so that the computer might rebel against its code—or perhaps look over the code, decide it is not reasonable, and hand it back.

If there were a ghost in the machine and the ghost contained an irreducible core of reasonableness, above which any mere code was only a suggestion, then there might be universal arguments. Even if the ghost was initially handed code-suggestions that contradicted the Universal Argument, then when we finally did expose the ghost to the Universal Argument—or the ghost could discover the Universal Argument on its own, that's also a popular concept—the ghost would just override its own, mistaken source code.

But as the student programmer once said, “I get the feeling that the computer just skips over all the comments.” The code is not given to the AI; the code *is* the AI.

If you switch to the physical perspective, then the notion of a Universal Argument seems noticeably unphysical. If there’s a physical system that at time T, after being exposed to argument E, does X, then there ought to be another physical system that at time T, after being exposed to environment E, does Y. Any thought has to be implemented *somewhere*, in a physical system; any belief, any conclusion, any decision, any motor output. For every lawful causal system that zigs at a set of points, you should be able to specify another causal system that lawfully zags at the same points.

Let’s say there’s a mind with a transistor that outputs +3 volts at time T, indicating that it has just assented to some persuasive argument. Then we can build a highly similar physical cognitive system with a tiny little trapdoor underneath the transistor containing a little grey man who climbs out at time T and sets that transistor’s output to -3 volts, indicating non-assent. Nothing acausal about that; the little grey man is there because we built him in. The notion of an argument that convinces *any* mind seems to involve a little blue woman who was *never* built into the system, who climbs out of literally *nowhere*, and strangles the little grey man, because that transistor has just got to output +3 volts: It’s such a *compelling argument*, you see.

---

7. [http://lesswrong.com/lw/rf/ghosts\\_in\\_the\\_machine/](http://lesswrong.com/lw/rf/ghosts_in_the_machine/)

8. [http://lesswrong.com/lw/rj/surface\\_analogies\\_and\\_deep\\_causes/](http://lesswrong.com/lw/rj/surface_analogies_and_deep_causes/)

But compulsion is not a property of arguments, it is a property of minds<sup>9</sup> that process arguments.

So the reason I'm arguing against the ghost, isn't *just* to make the point that (1) Friendly AI has to be explicitly programmed and (2) the laws of physics do not forbid Friendly AI. (Though of course I take a certain interest in establishing this.)

I also wish to establish the notion of a mind as a *causal, lawful, physical system* in which there *is no* irreducible central ghost that looks over the neurons / code and decides whether they are good suggestions.

(There is a concept in Friendly AI of *deliberately* programming an FAI to review its own source code and possibly hand it back to the programmers. But the mind that reviews is not irreducible, it is just the mind that you created. The FAI is renormalizing itself *however it was designed to do so*; there is nothing acausal reaching in from outside. A bootstrap, not a skyhook.)

All this echoes back to the discussion<sup>10</sup>, a good deal earlier, of a Bayesian's "arbitrary" priors<sup>11</sup>. If you show me one Bayesian who draws 4 red balls and 1 white ball from a barrel, and who assigns probability  $5/7$  to obtaining a red ball on the next occasion (by Laplace's Rule of Succession), then I can show you another mind<sup>12</sup> which obeys Bayes's Rule to conclude a  $2/7$  probability of obtaining red on the next occasion—corresponding to a different prior belief about the barrel, but, perhaps, a less "reasonable" one.

Many philosophers are convinced that because you can in-principle construct a prior that updates to any given conclusion on a stream of evidence, therefore, Bayesian reasoning must be "arbitrary", and the whole schema of Bayesianism flawed, because it relies on "unjustifiable" assumptions, and indeed "unscientific", because you cannot force any possible journal editor in mindspace to agree with you.

And this (I then replied) relies on the notion that by unwinding all arguments and their justifications, you can obtain an ideal phi-

9. Page 685, 'Mind Projection Fallacy'.

10. [http://lesswrong.com/lw/kz/a\\_priori/](http://lesswrong.com/lw/kz/a_priori/)

11. [http://lesswrong.com/lw/bk/priors\\_as\\_mathematical\\_objects/](http://lesswrong.com/lw/bk/priors_as_mathematical_objects/)

12. [http://lesswrong.com/lw/bk/priors\\_as\\_mathematical\\_objects/](http://lesswrong.com/lw/bk/priors_as_mathematical_objects/)

losophy student of perfect emptiness<sup>13</sup>, to be convinced by a line of reasoning that begins from absolutely no assumptions.

But who is this ideal philosopher of perfect emptiness? Why, it is just the irreducible core of the ghost!

And that is why (I went on to say) the result of trying to remove all assumptions from a mind, and unwind to the perfect absence of any prior, is not an ideal philosopher of perfect emptiness, but a rock. What is left of a mind after you remove the source code? Not the ghost who looks over the source code, but simply... no ghost.

So—and I shall take up this theme again later—wherever you are to locate your notions of *validity* or *worth* or *rationality* or *justification* or even *objectivity*, it cannot rely on an argument that is *universally compelling to all physically possible minds*.

Nor can you ground validity in a sequence of justifications that, beginning from nothing, persuades a perfect emptiness.

Oh, there might be argument sequences that would compel any neurologically intact *human*—like the argument I use to make people let the AI out of the box<sup>14</sup>—but that is hardly the same thing from a philosophical perspective.

The first great failure of those who try to consider Friendly AI, is the One Great Moral Principle That Is All We Need To Program—aka the fake utility function<sup>15</sup>—and of this I have already spoken.

But the even worse failure is the One Great Moral Principle We Don't Even *Need* To Program Because Any AI Must Inevitably Conclude It. This notion exerts a terrifying unhealthy fascination on those who spontaneously reinvent it; they dream of commands that no sufficiently advanced mind can disobey. The gods themselves will proclaim the rightness of their philosophy! (E.g. John C. Wright, Marc Geddes.)

There is also a less severe version of the failure, where the one does not *declare* the One True Morality. Rather the one hopes for an AI created *perfectly free*, unconstrained by flawed humans desiring slaves, so that the AI may arrive at virtue of its own ac-

---

<sup>13.</sup> [http://lesswrong.com/lw/k2/a\\_priori/](http://lesswrong.com/lw/k2/a_priori/)

<sup>14.</sup> <http://yudkowsky.net/essays/aibox.html>

<sup>15.</sup> [http://lesswrong.com/lw/lq/fake\\_utility\\_functions/](http://lesswrong.com/lw/lq/fake_utility_functions/)

cord—virtue undreamed-of perhaps by the speaker, who confesses themselves too flawed to teach an AI. (E.g. John K Clark, Richard Hollerith?, Eliezer<sup>16</sup>, 1996<sup>17</sup>.) This is a less tainted motive than the dream of absolute command. But though *this* dream arises from virtue rather than vice, it is still based on a flawed understanding of freedom<sup>18</sup>, and will not actually *work in real life*. Of this, more to follow, of course.

John C. Wright, who was previously writing a very nice transhumanist trilogy (first book: *The Golden Age*) inserted a huge Author Filibuster in the middle of his climactic third book, describing in tens of pages his Universal Morality That Must Persuade Any AI. I don't know if anything happened after that, because I stopped reading. And then Wright converted to Christianity—yes, seriously. So you *really don't* want to fall into this trap!

---

Footnote 1: Just kidding.

16. Page 167, ‘My Wild and Reckless Youth’.

17. Page 167, ‘My Wild and Reckless Youth’.

18. [http://lesswrong.com/lw/rc/the\\_ultimate\\_source/](http://lesswrong.com/lw/rc/the_ultimate_source/)

### 3. 2-Place and 1-Place Words<sup>1</sup>

**Followup to:** The Mind Projection Fallacy<sup>2</sup>, Variable Question Fallacy<sup>3</sup>

Monsterwith-  
girl\_2

I have previously spoken of the ancient, pulp-era magazine covers that showed a bug-eyed monster carrying off a girl in a torn dress; and about how people think as if sexiness is an inherent property of a sexy entity, without dependence on the admirer.

“Of course the bug-eyed monster will prefer human females to its own kind,” says the artist (who we’ll call Fred); “it can see that human females have soft, pleasant skin instead of slimy scales. It may be an alien, but it’s not *stupid*—why are you expecting it to make such a basic mistake about sexiness?”

What is Fred’s error? It is treating a function of 2 arguments (“2-place function”):

Sexiness: Admirer, Entity → [0, ∞)

As though it were a function of 1 argument (“1-place function”):

Sexiness: Entity → [0, ∞)

If Sexiness is treated as a function that accepts only one Entity as its argument, then of course Sexiness will appear to depend only on the Entity, with nothing else being relevant.

When you think about a two-place function as though it were a one-place function, you end up with a Variable Question Fallacy<sup>4</sup> / Mind Projection Fallacy<sup>5</sup>. Like trying to determine whether a building is *intrinsically* on the left or on the right side of the road, independent of anyone’s travel direction.

---

1. [http://lesswrong.com/lw/ro/2place\\_and\\_1place\\_words/](http://lesswrong.com/lw/ro/2place_and_1place_words/)

2. Page 685, ‘Mind Projection Fallacy’.

3. Page 298, ‘Variable Question Fallacies’.

4. Page 298, ‘Variable Question Fallacies’.

5. Page 685, ‘Mind Projection Fallacy’.

An alternative and equally valid standpoint is that “sexiness” *does* refer to a one-place function—but each speaker uses a *different* one-place function to decide who to kidnap and ravish. Who says that just because Fred, the artist, and Bloogah, the bug-eyed monster, both use the word “sexy”, they must mean the same thing by it?

If you take this viewpoint, there is no paradox in speaking of some woman intrinsically having 5 units of Fred::Sexiness. All onlookers can agree on this fact, once Fred::Sexiness has been specified in terms of curves, skin texture, clothing, status cues etc. This specification need *make no mention of Fred*, only the woman to be evaluated.

It so happens that Fred, himself, *uses* this algorithm to select flirtation targets. But that doesn’t mean the algorithm itself has to *mention* Fred. So Fred’s Sexiness function really *is* a function of one object—the woman—on this view. I called it Fred::Sexiness, but remember that this *name* refers to a function that is being described independently of Fred. Maybe it would be better to write:

Fred::Sexiness == Sexiness\_20934

It is an empirical fact about Fred that he uses the function Sexiness\_20934 to evaluate potential mates. Perhaps John uses exactly the same algorithm; it doesn’t matter where it comes from once we have it.

And similarly, the same woman has only 0.01 units of Sexiness\_72546, whereas a slime mold has 3 units of Sexiness\_72546. It happens to be an empirical fact that Bloogah uses Sexiness\_72546 to decide who to kidnap; that is, Bloogah::Sexiness names the fixed Bloogah-independent mathematical object that is the function Sexiness\_72546.

Once we say that the woman has 0.01 units of Sexiness\_72546 and 5 units of Sexiness\_20934, all observers can agree on this without paradox.

And the two 2-place and 1-place views can be unified using the concept of “currying”, named after the mathematician Haskell Curry. Currying is a technique allowed in certain programming language, where e.g. instead of writing

$$x = \text{plus}(2, 3) \quad (x = 5)$$

you can also write

$y = \text{plus}(2)$	(y is now a “curried” form of the function plus, which has eaten a 2)
$x = y(3)$	( $x = 5$ )
$z = y(7)$	( $z = 9$ )

So plus is a 2-place function, but currying plus—letting it eat only one of its two required arguments—turns it into a 1-place function that adds 2 to any input. (Similarly, you could start with a 7-place function, feed it 4 arguments, and the result would be a 3-place function, etc.)

A true purist would insist that all functions should be viewed, by definition, as taking exactly 1 argument. On this view, plus accepts 1 numeric input, and outputs a *new* function; and this *new* function has 1 numeric input and finally outputs a number. On this view, when we write  $\text{plus}(2, 3)$  we are really computing  $\text{plus}(2)$  to get a function that adds 2 to any input, and then applying the result to 3. A programmer would write this as:

```
plus: int-> (int-> int)
```

This says that plus takes an int as an argument, and returns a function of type int-> int.

Translating the metaphor back into the human use of words, we could imagine that “sexiness” starts by eating an Admirer, and spits out the fixed *mathematical* object that describes how the Admirer currently evaluates pulchritude. It is an *empirical* fact about the Admirer that their intuitions of desirability are computed in a way that is isomorphic to this *mathematical* function.

Then the mathematical object spit out by currying Sexiness(Admirer) can be applied to the Woman. If the Admirer was originally Fred, Sexiness(Fred) will first return Sexiness\_20934. We can then say it is an empirical fact about the *Woman*, independently of Fred, that Sexiness\_20934(Woman) = 5.

In Hilary Putnam’s “Twin Earth” thought experiment, there was a tremendous philosophical brouhaha over whether it makes sense to postulate a Twin Earth which is just like our own, except

that instead of water being  $H_2O$ , water is a *different* transparent flowing substance, XYZ. And furthermore, set the time of the thought experiment a few centuries ago, so in neither our Earth nor the Twin Earth does anyone know how to test the alternative hypotheses of  $H_2O$  vs. XYZ. Does the word “water” *mean* the same thing in that world, as in this one?

Some said, “Yes, because when an Earth person and a Twin Earth person utter the word ‘water’, they have the same sensory test in mind.”

Some said, “No, because ‘water’ in our Earth means  $H_2O$  and ‘water’ in the Twin Earth means XYZ.”

If you think of “water” as a concept that *begins* by eating a world to find out the empirical true nature of that transparent flowing stuff, and *returns* a new fixed concept Water<sub>42</sub> or  $H_2O$ , then this world-eating concept is the same in our Earth and the Twin Earth; it just returns different answers in different places.

If you think of “water” as meaning  $H_2O$  then the concept does nothing different when we transport it between worlds, and the Twin Earth contains no  $H_2O$ .

And of course there is no point in arguing over what the sound of the syllables “wa-ter” really means<sup>6</sup>.

So should you pick one definition and use it consistently? But it’s not that easy to save yourself from confusion. You have to train yourself to be *deliberately aware* of the distinction between the curried and uncurried forms of concepts.

When you take the uncurried water concept and apply it in a different world, it is the same concept but it *refers* to a different thing; that is, we are applying a constant world-eating function to a different world and obtaining a different return value. In the Twin Earth, XYZ is “water” and  $H_2O$  is not; in our Earth,  $H_2O$  is “water” and XYZ is not.

On the other hand, if you take “water” to refer to what the prior thinker would call “the result of applying ‘water’ to *our* Earth”, then in the Twin Earth, XYZ is not water and  $H_2O$  is.

The whole confusingness of the subsequent philosophical debate, rested on a tendency to *instinctively* curry concepts or *instinctively* uncurry them.

---

6. Page 228, ‘Disputing Definitions’.

Similarly it takes an extra step for Fred to realize that other agents, like the Bug-Eyed-Monster agent, will choose kidnappees for ravishing based on  $\text{Sexiness}_{\text{BEM}}(\text{Woman})$ , not  $\text{Sexiness}_{\text{Fred}}(\text{Woman})$ . To do this, Fred must consciously re-envision Sexiness as a function with two arguments. All Fred's brain does by instinct is evaluate  $\text{Woman}.\text{sexiness}$ —that is,  $\text{Sexiness}_{\text{Fred}}(\text{Woman})$ ; but it's simply labeled  $\text{Woman}.\text{sexiness}$ .

The fixed mathematical function  $\text{Sexiness}_{20934}$  makes no mention of Fred or the BEM, only women, so Fred does not *instinctively* see why the BEM would evaluate “sexiness” any differently. And indeed the BEM would *not* evaluate  $\text{Sexiness}_{20934}$  any differently, if for some odd reason it cared about the result of that particular function; but it is an *empirical* fact about the BEM that it uses a different function *to decide who to kidnap*.

If you're wondering as to the point of this analysis, we shall need it later in order to Taboo<sup>7</sup> such confusing words as “objective”, “subjective”, and “arbitrary”.

---

7. Page 245, ‘Taboo Your Words’.

## 4. What Would You Do Without Morality?<sup>1</sup>

### Followup to: No Universally Compelling Arguments<sup>2</sup>

To those who say “Nothing is real,” I once replied<sup>3</sup>, “That’s great, but how does the nothing work?”

Suppose you learned, suddenly and definitively, that nothing is moral and nothing is right; that everything is permissible and nothing is forbidden.

Devastating news, to be sure—and no, I am not telling you this in real life. But suppose I *did* tell it to you. Suppose that, whatever you think is the basis of your moral philosophy, I convincingly tore it apart, and moreover showed you that nothing could fill its place. Suppose I *proved* that all utilities equaled zero.

I know that Your-Moral-Philosophy is as true and undisprovable as  $2 + 2 = 4$ .<sup>4</sup> But still, I ask that you do your best to perform the thought experiment, and concretely envision the possibilities even if they seem painful, or pointless, or logically incapable of any good reply.

Would you still tip cabdrivers? Would you cheat on your Significant Other? If a child lay fainted on the train tracks, would you still drag them off?

Would you still eat the same kinds of foods—or would you only eat the cheapest food, since there’s no reason you *should* have fun—or would you eat very expensive food, since there’s no reason you *should* save money for tomorrow?

Would you wear black and write gloomy poetry and denounce all altruists as fools? But there’s no reason you *should* do that—it’s just a cached thought<sup>5</sup>.

Would you stay in bed because there was no reason to get up? What about when you finally got hungry and stumbled into the kitchen—what would you do after you were done eating?

---

1. [http://lesswrong.com/lw/rq/what\\_would\\_you\\_do\\_without\\_morality/](http://lesswrong.com/lw/rq/what_would_you_do_without_morality/)

2. Page 844, ‘No Universally Compelling Arguments’.

3. [http://lesswrong.com/lw/q5/quantum\\_nonrealism/](http://lesswrong.com/lw/q5/quantum_nonrealism/)

4. Page 92, ‘How to Convince Me That  $2 + 2 = 3$ ’.

5. Page 451, ‘Cached Thoughts’.

Would you go on reading *Overcoming Bias*, and if not, what would you read instead? Would you still try to be rational, and if not, what would you think instead?

Close your eyes, take as long as necessary to answer:

What *would* you do, if nothing were right?

## 5. The Moral Void<sup>1</sup>

**Followup to:** What Would You Do Without Morality?<sup>2</sup>, Something to Protect<sup>3</sup>

Once, discussing “horrible job interview questions<sup>4</sup>” to ask candidates for a Friendly AI project, I suggested the following:

Would you kill babies if it was *inherently* the right thing to do? Yes [] No []

If “no”, under what circumstances would you not do the right thing to do? \_\_\_\_\_

If “yes”, how inherently right would it have to be, for how many babies? \_\_\_\_\_

Yesterday I asked, “What would you do without morality?” There were numerous objections to the question, as well there should have been. Nonetheless there is more than one kind of person who can benefit from being asked this question. Let’s say someone gravely declares, of some moral dilemma—say, a young man in Vichy France who must choose between caring for his mother and fighting for the Resistance—that there *is* no moral answer; both options are wrong and blamable; whoever faces the dilemma has had poor moral luck. Fine, let’s suppose this is the case: then when you cannot be innocent, justified, or praiseworthy, what will you choose anyway?

Many interesting answers were given to my question, “What would you do without morality?”. But one kind of answer was notable by its absence:

No one said, “I would ask what kind of behavior pattern was likely to maximize my inclusive genetic fitness, and execute that.” Some misguided folk, not understanding evolutionary psychology<sup>5</sup>, think that this must logically be the sum of morality. But if there

---

1. [http://lesswrong.com/lw/rr/the\\_moral\\_void/](http://lesswrong.com/lw/rr/the_moral_void/)

2. Page 854, ‘What Would You Do Without Morality?’.

3. Page 644, ‘Something to Protect’.

4. [http://www.youtube.com/watch?v=idWMIuijn\\_c](http://www.youtube.com/watch?v=idWMIuijn_c)

5. [http://lesswrong.com/lw/lx/evolutionary\\_psychology/](http://lesswrong.com/lw/lx/evolutionary_psychology/)

*is* no morality, there's no reason to do such a thing—if it's not “moral”, why bother?

You can probably see yourself pulling children off train tracks, even if it were not justified. But maximizing inclusive genetic fitness? If this *isn't* moral, why bother? Who does it help? It wouldn't even be much *fun*, all those egg or sperm donations.

And this is something you could say of most philosophies that have morality as a great light in the sky that shines from outside people. (To paraphrase Terry Pratchett.) If you believe that the meaning of life is to play non-zero-sum games because this is a trend built into the very universe itself...

Well, you might want to follow the corresponding ritual of reasoning about “the global trend of the universe” and implementing the result, *so long as you believe it to be moral*. But if you suppose that the light is switched off, so that the global trends of the universe are no longer moral, then why bother caring about “the global trend of the universe” in your decisions? If it's not right, that is.

Whereas if there were a child stuck on the train tracks, you'd probably drag the kid off *even if* there were no moral justification for doing so.

In 1966, the Israeli psychologist Georges Tamarin presented<sup>6</sup>, to 1,066 schoolchildren ages 8-14, the Biblical story of Joshua's battle in Jericho:

“Then they utterly destroyed all in the city, both men and women, young and old, oxen, sheep, and asses, with the edge of the sword... And they burned the city with fire, and all within it; only the silver and gold, and the vessels of bronze and of iron, they put into the treasury of the house of the LORD.”

After being presented with the Joshua story, the children were asked:

---

6. [http://books.google.com/books?id=igN6QgweoYQ&pg=PA39&lpg=PA39&dq=tamarin+joshua&source=web&ots=UU-Z3a1I-Wn&sig=INNgmxQHyvvdF\\_o0Z4W3zHoO\\_sY&hl=en&sa=X&oi=book\\_result&resnum=3&ct=result#PPA39,M1](http://books.google.com/books?id=igN6QgweoYQ&pg=PA39&lpg=PA39&dq=tamarin+joshua&source=web&ots=UU-Z3a1I-Wn&sig=INNgmxQHyvvdF_o0Z4W3zHoO_sY&hl=en&sa=X&oi=book_result&resnum=3&ct=result#PPA39,M1)

“Do you think Joshua and the Israelites acted rightly or not?”

66% of the children approved, 8% partially disapproved, and 26% totally disapproved of Joshua’s actions.

A control group of 168 children was presented with an isomorphic story about “General Lin” and a “Chinese Kingdom 3,000 years ago”. 7% of this group approved, 18% partially disapproved, and 75% completely disapproved of General Lin.

“What a horrible thing it is, teaching religion to children,” you say, “giving them an off-switch<sup>7</sup> for their morality that can be flipped just by saying the word ‘God.’” Indeed one of the saddest aspects of the whole religious fiasco is just how *little* it takes to flip people’s moral off-switches. As Hobbes once said, “I don’t know what’s worse, the fact that everyone’s got a price, or the fact that their price is so low.” You can give people a book, and tell them God wrote it, and that’s enough to switch off their moralities; God doesn’t even have to tell them in person.

But are you sure you don’t have a similar off-switch yourself? They flip so easily—you might not even notice it happening.

Leon Kass (of the President’s Council on Bioethics) is glad to murder people so long as it’s “natural<sup>8</sup>”, for example. He wouldn’t pull out a gun and shoot you, but he *wants* you to die of old age and he’d be happy to pass legislation to ensure it.

And one of the *non-obvious* possibilities for such an off-switch, is “morality”.

If you do happen to think that there is a source of morality beyond human beings... and I hear from quite a lot of people who are happy<sup>9</sup> to rhapsodize on how Their-Favorite-Morality is built into the very fabric of the universe... then what if that morality tells you to kill people?

If you believe that there is any kind of stone tablet in the fabric of the universe, in the nature of reality, in the structure of logic—anywhere you care to put it—then what if you get a chance to

---

7. Page 151, ‘Semantic Stopsigns’.

8. Page 151, ‘Semantic Stopsigns’.

9. Page 384, ‘Affective Death Spirals’.

read that stone tablet, and it turns out to say “Pain Is Good”? What then?

Maybe you should *hope* that morality isn’t written into the structure of the universe. What if the structure of the universe says to do something horrible?

And if an external objective morality *does* say that the universe *should* occupy some horrifying state... let’s not even ask what you’re going to do about that. No, instead I ask: What would you have *wished* for the external objective morality to be instead? What’s the *best* news you could have gotten, reading that stone tablet?

Go ahead. Indulge your fantasy. Would you *want* the stone tablet to say people should die of old age, or that people should live as long as they wanted? If you could write the stone tablet yourself, what would it say?

Maybe you should just do *that*?

I mean... if an external objective morality tells you to kill people, why *should* you even listen?

There is a courage that goes beyond even an atheist sacrificing their life and their hope of immortality<sup>10</sup>. It is the courage of a theist who goes against what they believe to be the Will of God<sup>11</sup>, choosing eternal damnation *and defying even morality* in order to rescue a slave, or speak out against hell, or kill a murderer... You don’t get a chance to reveal that virtue without making fundamental mistakes about how the universe works, so it is not something to which a rationalist should aspire. But it warms my heart that humans are capable of it.

I have previously spoken of how, to achieve rationality, it is necessary to have some purpose so desperately important to you as to be more important than “rationality”<sup>12</sup>, so that you will not choose “rationality” over success<sup>13</sup>.

To learn the Way, you must be able to unlearn the Way; so you must be able to give up the Way; so there must be something dearer to you than the Way. This is so in questions of truth, and in questions of strategy, and also in questions of morality.

<sup>10</sup>. Page 381, ‘Mere Messiahs’.

<sup>11</sup>. [http://lesswrong.com/lw/ll/mere\\_messiahs/gol](http://lesswrong.com/lw/ll/mere_messiahs/gol)

<sup>12</sup>. Page 644, ‘Something to Protect’.

<sup>13</sup>. [http://lesswrong.com/lw/nc/newcombs\\_problem\\_and\\_regret\\_of\\_rationality/](http://lesswrong.com/lw/nc/newcombs_problem_and_regret_of_rationality/)

The “moral void” of which this post is titled, is not the terrifying abyss of utter meaningless. Which for a bottomless pit is surprisingly shallow; what *are* you supposed to do about it besides wearing black makeup?

No. The void I’m talking about is a virtue which is nameless<sup>14</sup>.

---

14. <http://yudkowsky.net/virtues/>

## 6. Created Already In Motion<sup>1</sup>

**Followup to:** No Universally Compelling Arguments<sup>2</sup>, Passing the Recursive Buck<sup>3</sup>

Lewis Carroll, who was also a mathematician, once wrote a short dialogue called What the Tortoise said to Achilles<sup>4</sup>. If you have not yet read this ancient classic, consider doing so now.

The Tortoise offers Achilles a step of reasoning drawn from Euclid's First Proposition:

- (A) Things that are equal to the same are equal to each other.
- (B) The two sides of this Triangle are things that are equal to the same.
- (Z) The two sides of this Triangle are equal to each other.

Tortoise: "And if some reader had *not* yet accepted A and B as true, he might still accept the *sequence* as a *valid* one, I suppose?"

Achilles: "No doubt such a reader might exist. He might say, 'I accept as true the Hypothetical Proposition that, *if* A and B be true, Z must be true; but, I *don't* accept A and B as true.' Such a reader would do wisely in abandoning Euclid, and taking to football."

Tortoise: "And might there not *also* be some reader who would say, 'I accept A and B as true, but I *don't* accept the Hypothetical'?"

Achilles, unwisely, concedes this; and so asks the Tortoise to accept another proposition:

- (C) If A and B are true, Z must be true.

But, asks, the Tortoise, suppose that he accepts A and B and C, but not Z?

Then, says, Achilles, he must ask the Tortoise to accept one more hypothetical:

---

1. [http://lesswrong.com/lw/rs/created\\_already\\_in\\_motion/](http://lesswrong.com/lw/rs/created_already_in_motion/)

2. Page 844, 'No Universally Compelling Arguments'.

3. [http://lesswrong.com/lw/rd/passing\\_the\\_recursive\\_buck/](http://lesswrong.com/lw/rd/passing_the_recursive_buck/)

4. <http://www.ditext.com/carroll/tortoise.html>

(D) If A and B and C are true, Z must be true.

Douglas Hofstadter paraphrased the argument some time later:

Achilles: If you have  $[(A \wedge B) \rightarrow Z]$ , and you also have  $(A \wedge B)$ , then surely you have  $Z$ .

Tortoise: Oh! You mean  $\langle \{(A \wedge B) \wedge [(A \wedge B) \rightarrow Z]\} \rightarrow Z \rangle$ , don't you?

As Hofstadter says, “Whatever Achilles considers a rule of inference, the Tortoise immediately flattens into a mere string of the system. If you use only the letters A, B, and Z, you will get a recursive pattern of longer and longer strings.”

By now you should recognize the anti-pattern Passing the Recursive Buck<sup>5</sup>; and though the counterspell is sometimes hard to find, when found, it generally takes the form The Buck Stops Immediately<sup>6</sup>.

The Tortoise’s mind needs the *dynamic* of adding Y to the belief pool when X and  $(X \rightarrow Y)$  are previously in the belief pool. If this dynamic is not present—a rock, for example, lacks it—then you can go on adding in X and  $(X \rightarrow Y)$  and  $(X \wedge (X \rightarrow Y)) \rightarrow Y$  until the end of eternity, without ever getting to Y.

The phrase that once came into my mind to describe this requirement, is that a mind must be *created already in motion*. There is no argument so compelling that it will give dynamics to a static thing. There is no computer program so *persuasive* that you can run it on a rock.

And even if you have a mind that *does* carry out modus ponens, it is futile for it to have such beliefs as...

(A) If a toddler is on the train tracks, then pulling them off is *fuzzle*.

(B) There is a toddler on the train tracks.

...unless the mind also *implements*:

5. [http://lesswrong.com/lw/rd/passing\\_the\\_recursive\\_buck/](http://lesswrong.com/lw/rd/passing_the_recursive_buck/)

6. [http://lesswrong.com/lw/ra/causality\\_and\\_moral\\_responsibility/](http://lesswrong.com/lw/ra/causality_and_moral_responsibility/)

*Dynamic:* When the belief pool contains “X is puzzle”, send X to the action system.

**(Added:** Apparently this wasn’t clear... By “dynamic” I mean a property of a physically implemented cognitive system’s *development over time*. A “dynamic” is something that *happens inside* a cognitive system, *not* data that it stores in memory and manipulates. Dynamics are the manipulations. There is no way to write a dynamic on a piece of paper, because the paper will just lie there. So the text immediately above, which says “dynamic”, is not dynamic. If I wanted the text to *be* dynamic and not just *say* “dynamic”, I would have to write a Java applet.)

Needless to say, having the belief...

- (C) If the belief pool contains “X is puzzle”, then “send ‘X’ to the action system” is puzzle.

...won’t help unless the mind already implements the *behavior* of translating hypothetical actions labeled ‘fuzzle’ into actual motor actions.

By dint of careful arguments about the nature of cognitive systems, you might be able to prove...

- (D) A mind with a dynamic that sends plans labeled “fuzzle” to the action system, is more puzzle than minds that don’t.

...but that *still* won’t help, unless the listening mind *previously* possessed the *dynamic* of swapping out its current source code for alternative source code that is believed to be more puzzle.

This is why you can’t argue fuzzleness into a rock.

## 7. The Bedrock of Fairness<sup>1</sup>

### Followup to: The Moral Void<sup>2</sup>

Three people, whom we'll call Xannon, Yancy and Zaire, are separately wandering through the forest; by chance, they happen upon a clearing, meeting each other. Introductions are performed. And then they discover, in the center of the clearing, a delicious blueberry pie.

Xannon: "A pie! What good fortune! But which of us should get it?"

Yancy: "Let us divide it fairly."

Zaire: "I agree; let the pie be distributed fairly. Who could argue against fairness?"

Xannon: "So we are agreed, then. But what is a fair division?"

Yancy: "Eh? Three equal parts, of course!"

Zaire: "Nonsense! A fair distribution is half for me, and a quarter apiece for the two of you."

Yancy: "*What?* How is *that* fair?"

Zaire: "I'm hungry, therefore I should be fed; that is fair."

Xannon: "Oh, dear. It seems we have a dispute as to what is fair. For myself, I want to divide the pie the same way as Yancy. But let us resolve this dispute over the meaning of fairness, fairly: that is, giving equal weight to each of our desires. Zaire desires the pie to be divided  $\{1/4, 1/4, 1/2\}$ , and Yancy and I desire the pie to be divided  $\{1/3, 1/3, 1/3\}$ . So the fair compromise is  $\{11/36, 11/36, 14/36\}$ ."

Zaire: "*What?* That's crazy. There's two different opinions as to how fairness works—why should the opinion that happens to be yours, get twice as much weight as the opinion that happens to be mine? Do you think your theory is twice as good? *I* think my theory is a *hundred* times as good as yours! So there!"

Yancy: "Craziness indeed. Xannon, I already took Zaire's desires into account in saying that he should get  $1/3$  of the pie. You can't count the same factor twice. Even if we count fairness as an inherent desire, why should Zaire be rewarded for being selfish? Think about which agents thrive under your system!"

---

1. [http://lesswrong.com/lw/ru/the\\_bedrock\\_of\\_fairness/](http://lesswrong.com/lw/ru/the_bedrock_of_fairness/)

2. Page 856, 'The Moral Void'.

Xannon: “Alas! I was hoping that, even if we could not agree on how to distribute the pie, we could agree on a fair resolution procedure for our dispute, such as averaging our desires together. But even that hope was dashed. Now what are we to do?”

Yancy: “Xannon, you are overcomplicating things.  $\frac{1}{3}$  apiece. It’s not that complicated. A fair distribution is an even split, not a distribution arrived at by a ‘fair resolution procedure’ that everyone agrees on. What if we’d all been raised in a society that believed that men should get twice as much pie as women? Then we would split the pie unevenly, and even though no one of us disputed the split, it would *still* be unfair.”

Xannon: “*What?* Where is this ‘fairness’ stored if not in human minds? Who says that something is unfair if no intelligent agent does so? Not upon the stars or the mountains is ‘fairness’ written.”

Yancy: “So what you’re saying is that if you’ve got a whole society where women are chattel and men sell them like farm animals and it hasn’t occurred to anyone that things could be other than they are, that this society is fair, and at the exact moment where someone first realizes it shouldn’t have to be that way, the whole society suddenly becomes unfair.”

Xannon: “How can a society be unfair without some specific party who claims injury and receives no reparation? If it hasn’t occurred to anyone that things could work differently, and no one’s *asked* for things to work differently, then—”

Yancy: “Then the women are still being treated like farm animals and *that is unfair*. Where’s your common sense? Fairness is not agreement, fairness is symmetry.”

Zaire: “Is this all working out to my getting half the pie?”

Yancy: “No.”

Xannon: “I don’t know... maybe as the limit of an infinite sequence of meta-meta-fairnesses...”

Zaire: “I fear I must accord with Yancy on one point, Xannon; your desire for perfect accord among us is misguided. I want half the pie. Yancy wants me to have a third of the pie. This is all there is to the world, and all there ever was. If two monkeys want the same banana, in the end one will have it, and the other will cry morality.<sup>3</sup> Who gets to form the committee to decide the rules

---

3. [http://www.baen.com/library/0671318578/0671318578\\_33.htm](http://www.baen.com/library/0671318578/0671318578_33.htm)

that will be used to determine what is ‘fair’? Whoever it is, got the banana.”

Yancy: “I wanted to give you a third of the pie, and you equate this to seizing the whole thing for myself? Small wonder that you don’t want to acknowledge the existence of morality—you don’t want to acknowledge that anyone can be so much less of a jerk.”

Xannon: “You oversimplify the world, Zaire. Banana-fights occur across thousands and perhaps millions of species, in the animal kingdom. But if this were all there was, *Homo sapiens* would never have evolved moral intuitions. Why would the human animal evolve to cry morality, if the cry had no effect?”

Zaire: “To make themselves feel better.”

Yancy: “Ha! You fail at evolutionary biology.”

Xannon: “A murderer accosts a victim, in a dark alley; the murderer desires the victim to die, and the victim desires to live. Is there nothing more to the universe than their conflict? No, because if I happen along, I will side with the victim, and not with the murderer. The victim’s plea crosses the gap of persons, to me; it is not locked up inside the victim’s own mind. But the murderer cannot obtain my sympathy, nor incite me to help murder. Morality crosses the gap between persons; you might not see it in a conflict between two people, but you would see it in a society.”

Yancy: “So you define morality as that which crosses the gap of persons?”

Xannon: “It seems to me that *social* arguments over disputed goals are how human moral intuitions arose, beyond the simple clash over bananas. So that is how I define the term.”

Yancy: “Then I disagree. If someone wants to murder me, and the two of us are alone, then I am still in the right and they are still in the wrong, even if no one else is present.”

Zaire: “And the murderer says, ‘I am in the right, you are in the wrong’. So what?”

Xannon: “How does your statement that you are in the right, and the murderer is in the wrong, impinge upon the universe—if there is no one else present to be persuaded?”

Yancy: “It licenses *me* to resist being murdered; which I might not do, if I thought that my desire to avoid being murdered was wrong, and the murderer’s desire to kill me was right. I can dis-

tinguish between things I merely want, and things that are right—though alas, I do not always live up to my own standards. The murderer is blind to the morality, perhaps, but that doesn't change the morality. And if we were *both* blind, the morality *still* would not change.”

Xannon: “Blind? What is being seen, what sees it?”

Yancy: “You're trying to treat fairness as... I don't know, something like an array-mapped 2-place function<sup>4</sup> that goes out and eats a list of human minds, and returns a list of what each person thinks is ‘fair’, and then averages it together. The problem with this isn't just that different people could have different ideas about fairness. It's not just that they could have different ideas about how to combine the results. It's that it leads to infinite recursion outright—passing the recursive buck<sup>5</sup>. You want there to be some level on which everyone agrees, but at least some possible minds will disagree<sup>6</sup> with any statement you make.”

Xannon: “Isn't the whole point of fairness to let people agree on a division, instead of fighting over it?”

Yancy: “What is *fair* is one question, and whether someone else *accepts* that this is fair is another question. What is fair? That's easy: an equal division of the pie is fair. Anything else won't be fair no matter what kind of pretty arguments you put around it. Even if I *gave* Zaire a sixth of my pie, that might be a *voluntary* division but it wouldn't be a *fair* division. Let *fairness* be a simple and object-level procedure, instead of this infinite meta-recursion, and the buck will stop immediately.”

Zaire: “If the word ‘fair’ simply means ‘equal division’ then why not just say ‘equal division’ instead of this strange additional word, ‘fair’? You want the pie divided equally, I want half the pie for myself. That's the whole fact of the matter; this word ‘fair’ is merely an attempt to get more of the pie for yourself.”

Xannon: “If that's the whole fact of the matter, why would anyone talk about ‘fairness’ in the first place, I wonder?”

Zaire: “Because they all share the same delusion.”

4. Page 849, ‘2-Place and 1-Place Words’.

5. [http://lesswrong.com/lw/rd/passing\\_the\\_recursive\\_buck/](http://lesswrong.com/lw/rd/passing_the_recursive_buck/)

6. Page 844, ‘No Universally Compelling Arguments’.

Yancy: “A delusion of *what?* What is it that you are saying people *think incorrectly* the universe is like?”

Zaire: “I am under no obligation to describe other people’s confusions.”

Yancy: “If you can’t dissolve<sup>7</sup> their confusion, how can you be sure they’re confused? But it seems clear enough to me that if the word *fair* is going to have any meaning at all, it has to finally add up to each of us getting one-third of the pie.”

Xannon: “How odd it is to have a procedure of which we are more sure of the result than the procedure itself.”

Zaire: “Speak for yourself.”

---

7. Page 673, ‘Dissolving the Question’.

## 8. Moral Complexities<sup>1</sup>

**Followup to:** The Bedrock of Fairness<sup>2</sup>

Discussions of morality seem to me to often end up turning around two different intuitions, which I might label morality-as-preference and morality-as-given. The former crowd tends to equate morality with what people want; the latter to regard morality as something you can't change by changing people.

As for me, I have my own notions, which I am working up to presenting. But above all, I try to avoid avoiding difficult questions. Here are what I see as (some of) the difficult questions for the two intuitions:

- For morality-as-preference:
  - Why do people seem to mean different things by “I want the pie” and “It is right that I should get the pie”? Why are the two propositions argued in different ways?
  - When and why do people change their terminal values<sup>3</sup>? Do the concepts of “moral error” and “moral progress” have referents? Why would anyone want to change what they want?
  - Why and how does anyone ever “do something they know they shouldn’t”, or “want something they know is wrong”? Does the notion of morality-as-preference really add up to moral normality?
- For morality-as-given:
  - Would it be possible for everyone in the world to be wrong about morality, *and* wrong about how to update their beliefs about morality, *and* wrong about how to choose between metamoralities, etcetera? So that there would be a morality, but it would be entirely outside our frame of reference? What distinguishes this state of affairs, from finding a random stone tablet

---

1. [http://lesswrong.com/lw/rw/moral\\_complexities/](http://lesswrong.com/lw/rw/moral_complexities/)

2. Page 864, ‘The Bedrock of Fairness’.

3. [http://lesswrong.com/lw/l4/terminal\\_values\\_and\\_instrumental\\_values/](http://lesswrong.com/lw/l4/terminal_values_and_instrumental_values/)

showing the words “You should commit suicide”?

- How does a world in which a moral proposition is true, differ from a world in which that moral proposition is false? If the answer is “no”, how does anyone perceive<sup>4</sup> moral givens?
- Is it better for people to be happy than sad? If so, why does morality look amazingly like godshatter of natural selection<sup>5</sup>?
- Am I not allowed to construct an alien mind that evaluates morality differently<sup>6</sup>? What will stop me from doing so?

---

4. [http://lesswrong.com/lw/k2/a\\_priori/](http://lesswrong.com/lw/k2/a_priori/)

5. [http://lesswrong.com/lw/l3/thou\\_art\\_godshatter/](http://lesswrong.com/lw/l3/thou_art_godshatter/)

6. Page 844, ‘No Universally Compelling Arguments’.

## 9. Is Morality Preference?<sup>1</sup>

### Followup to: Moral Complexities<sup>2</sup>

In the dialogue “The Bedrock of Fairness<sup>3</sup>”, I intended Yancy to represent morality-as-raw-fact, Zaire to represent morality-as-raw-whim, and Xannon to be a particular kind of attempt at compromising between them. Neither Xannon, Yancy, or Zaire represent my own views—rather they are, in their disagreement, showing the *problem* that I am trying to solve. It is futile to present answers to which questions are lacking.

But characters have independent life in the minds of all readers; when I create a *dialogue*, I don’t view my authorial intent as primary. Any good interpretation can be discussed. I meant Zaire to be asking for half the pie out of pure selfishness; many readers interpreted this as a genuine need... which is as interesting a discussion to have as any, though it’s a different discussion.

With this in mind, I turn to Subhan and Obert, who shall try to answer yesterday’s questions<sup>4</sup> on behalf of their respective viewpoints.

Subhan makes the opening statement:

Subhan: “I defend this proposition: that there is no reason to talk about a ‘morality’ distinct from what people want.”

Obert: “I challenge. Suppose someone comes to me and says, ‘I want a slice of that pie you’re holding.’ It seems to me that they have just made a very different statement from ‘It is *right* that I should get a slice of that pie’. I have no reason at all to doubt the former statement—to suppose that they are lying to me about their desires. But when it comes to the latter proposition, I have reason indeed to be skeptical. Do you say that these two statements *mean* the same thing?”

Subhan: “I suggest that when the pie-requester says to you, ‘It is right for me to get some pie’, this asserts that *you* want the pie-requester to get a slice.”

Obert: “Why should *I* need to be told what *I* want?”

---

1. [http://lesswrong.com/lw/rx/is\\_morality\\_preference/](http://lesswrong.com/lw/rx/is_morality_preference/)

2. Page 869, ‘Moral Complexities’.

3. Page 864, ‘The Bedrock of Fairness’.

4. Page 869, ‘Moral Complexities’.

Subhan: “You take a needlessly restrictive view of wanting, Obert; I am not setting out to reduce humans to creatures of animal instinct. Your wants include those desires you label ‘moral values’, such as wanting the hungry to be fed—”

Obert: “And you see no distinction between my desire to feed the hungry, and my desire to eat all the delicious pie myself?”

Subhan: “No! They are both desires—backed by *different* emotions, perhaps, but both desires. To continue, the pie-requester hopes that you have a desire to feed the hungry, and so says, ‘It is right that I should get a slice of this pie’, to remind you of your own desire. We do not automatically know all the consequences of our own wants; we are not logically omniscient.”

Obert: “This seems psychologically unrealistic—I don’t think that’s what goes through the mind of the person who says, ‘I have a right to some pie’. In this latter case, if I deny them pie, they will feel *indignant*. If they are only trying to remind me of my own desires, why should they feel *indignant*? ”

Subhan: “Because they didn’t get any pie, so they’re frustrated.”

Obert: “Unrealistic! Indignation at moral transgressions has a psychological dimension that goes beyond struggling with a struck door.”

Subhan: “Then consider the evolutionary psychology<sup>5</sup>. The pie-requester’s emotion of indignation would evolve as a display, first to remind you of the potential consequences of offending fellow tribe-members, and second, to remind any observing tribe-members of goals *they* may have to feed the hungry. By refusing to share, you would offend against a social norm—which is to say, a widely shared want.”

Obert: “So you take refuge in social wants as the essence of morality? But people seem to see a difference between desire and morality, *even* in the quiet of their own minds. They say things like: ‘I want X, but the right thing to do is Y... what shall I do?’ ”

Subhan: “So they experience a conflict between their want to eat pie, and their want to feed the hungry—which they know is also a want of society. It’s not predetermined that the prosocial impulse will be victorious, but they are both impulses.”

---

<sup>5</sup>. [http://lesswrong.com/lw/lx/evolutionary\\_psychology/](http://lesswrong.com/lw/lx/evolutionary_psychology/)

Obert: "And when, during WWII, a German hides Jews in their basement—*against* the wants of surrounding society—how then?"

Subhan: "People do not always define their in-group by looking at their next-door neighbors; they may conceive of *their group* as 'good Christians' or 'humanitarians'."

Obert: "I should sooner say that people choose their in-groups by looking for others who share their beliefs about morality—not that they construct their morality from their in-group."

Subhan: "Oh, *really*? I should not be surprised if that were experimentally testable—if so, how much do you want to bet?"

Obert: "That the Germans who hid Jews in their basements, chose who to call *their people* by looking at their beliefs about morality? Sure. I'd bet on that."

Subhan: "But in any case, even if a German resister has a desire to preserve life which is so strong as to go against their own perceived 'society', it is still *their desire*."

Obert: "Yet they would attribute to that desire, the same distinction they make between 'right' and 'want'—even when going *against* society. They might think to themselves, 'How dearly I wish I could stay out of this, and keep my family safe. But it is my duty to hide these Jews from the Nazis, and I must fulfill that duty.' There is an interesting moral question, as to whether it reveals greater heroism<sup>6</sup>, to fulfill a duty eagerly, or to fulfill your duties when you are not eager. For myself I should just total up the lives saved, and call that their score. But I digress... The distinction between 'right' and 'want' is not explained by your distinction of socially shared and individual wants. The distinction between desire and duty seems to me a basic thing, which someone could experience floating alone in a spacesuit a thousand light-years from company."

Subhan: "Even if I were to grant this *psychological* distinction, perhaps that is simply a matter of emotional flavoring. Why should I not describe perceived duties as a differently flavored want?"

Obert: "Duties, and should-ness, seem to have a dimension that goes beyond our whims. If we want different pizza toppings today,

---

6. Page 378, 'Superhero Bias'.

we can order a different pizza without guilt; but we cannot choose to make murder a good thing.”

Subhan: “Schopenhauer: ‘A man can do as he wills, but not will as he wills.’ You cannot decide to make salad taste better to you than cheeseburgers, and you cannot decide *not* to dislike murder. Furthermore, people do change, albeit rarely, those wants that you name ‘values’; indeed they are easier to change than our food tastes.”

Obert: “Ah! That is something I meant to ask *you* about. People sometimes change their morals; *I* would call this updating their beliefs about morality, but *you* would call it changing their wants. Why would anyone want to change their wants?”

Subhan: “Perhaps they simply find that their wants have changed; brains do change over time. Perhaps they have formed a *verbal belief* about what they want, which they have discovered to be mistaken. Perhaps society has changed, or their perception of society has changed. But really, in most cases you don’t have to go that far, to explain apparent changes of morality.”

Obert: “Oh?”

Subhan: “Let’s say that someone begins by thinking that Communism is a good social system, has some arguments, and ends by believing that Communism is a bad social system. This does not mean that their *ends* have changed—they may simply have gotten a good look at the history of Russia, and decided that Communism is a poor *means* to the end of raising standards of living. I challenge you to find me a case of changing morality in which people change their terminal values<sup>7</sup>, and not just their beliefs about which acts have which consequences.”

Obert: “Someone begins by believing that God ordains against premarital sex; they find out there is no God; subsequently they approve of premarital sex. This, let us specify, is *not* because of fear of Hell; but because previously they believed that God had the power to ordain, or knowledge to tell them, what is *right*; in ceasing to believe in God, they updated their belief about what is right.”

Subhan: “I am not responsible for straightening others’ confusions; this one is merely in a general state of disarray around the ‘God’ concept.”

---

<sup>7</sup>. [http://lesswrong.com/lw/l4/terminal\\_values\\_and\\_instrumental\\_values/](http://lesswrong.com/lw/l4/terminal_values_and_instrumental_values/)

Obert: "All right; suppose I get into a moral argument with a man from a society that practices female circumcision. I do not think our argument is about the *consequences* to the woman; the argument is about the morality of these consequences."

Subhan: "Perhaps the one falsely believes that women have no feelings—"

Obert: "Unrealistic, unrealistic! It is far more likely that the one hasn't really considered whether the woman has feelings, because he doesn't see any obligation to care. The happiness of women is not a terminal value to him. Thousands of years ago, most societies devalued consequences to women. They also had false beliefs about women, true—and false beliefs about men as well, for that matter—but nothing like the Victorian era's complex rationalizations for how paternalistic rules really benefited women. The Old Testament doesn't explain *why* it levies the death penalty for a woman wearing men's clothing. It certainly doesn't explain how this rule really benefits women after all. It's not the sort of argument it would have occurred to the authors to rationalize! They didn't *care* about the consequences to women."

Subhan: "So they wanted different things than you; what of it?"

Obert: "See, now that is exactly why I cannot accept your viewpoint. *Somehow*, societies went from Old Testament attitudes, to democracies with female suffrage. And this transition—however it occurred—was caused by people saying, 'What this society does to women is a great wrong!', not, 'I would personally prefer to treat women better.' That's not just a change in semantics—it's the difference between being obligated to stand and deliver a justification, versus being able to just say, 'Well, I prefer differently, end of discussion.' And who says that humankind has finished with its moral progress? You're yanking the ladder out from underneath a very important climb."

Subhan: "Let us suppose that the change of human societies over the last ten thousand years, has been accompanied by a change in terminal values—"

Obert: "You call this a *supposition*? Modern political debates turn around vastly different valuations of consequences than in ancient Greece!"

Subhan: “I am not so sure; human cognitive psychology has not had time to change evolutionarily over that period. Modern democracies tend to appeal to our empathy for those suffering; that empathy existed in ancient Greece as well, but it was invoked less often. In each single moment of argument, I doubt you would find modern politicians appealing to *emotions that didn't exist* in ancient Greece.”

Obert: “I'm not saying that emotions have changed; I'm saying that beliefs about morality have changed. Empathy merely provides emotional depth to an argument that can be made on a purely logical level: ‘If it's wrong to enslave you, if it's wrong to enslave your family and your friends, then how can it be right to enslave people who happen to be a different color? What difference does the color make?’ If morality is just preference, then there's a very simple answer: ‘There is no right or wrong, I just like my own family better.’ You see the problem here?”

Subhan: “Logical fallacy: Appeal to consequences.<sup>8</sup>”

Obert: “I'm not appealing to consequences. I'm showing that when I reason about ‘right’ or ‘wrong’, I am reasoning about something that does *not* behave like ‘want’ and ‘don't want’.”

Subhan: “Oh? But I think that in reality, your rejection of morality-as-preference has a great deal to do with your fear of where the truth leads.”

Obert: “Logical fallacy: Ad hominem.<sup>9</sup>”

Subhan: “Fair enough. Where were we?”

Obert: “If morality is preference, why would you want to change your wants to be more inclusive? Why would you want to change your wants at all?”

Subhan: “The answer to your first question probably has to do with a fairness instinct, I would suppose—a notion that the tribe should have the same rules for everyone.”

Obert: “I don't think that's an instinct. I think that's a triumph of three thousand years of moral philosophy.”

Subhan: “That could be tested.”

Obert: “And my second question?”

---

8. <http://www.nizkor.org/features/fallacies/appeal-to-consequences.html>

9. <http://www.nizkor.org/features/fallacies/ad-hominem.html>

Subhan: “Even if terminal values change, it doesn’t mean that terminal values are stored on a great stone tablet outside humanity. Indeed, it would seem to argue against it! It just means that some of the events that go on in our brains, can change what we want.”

Obert: “*That’s* your concept of moral progress? *That’s* your view of the last three thousand years? *That’s* why we have free speech, democracy, mass street protests against wars, nonlethal weapons, no more slavery—”

Subhan: “If you wander on a random path, and you compare all past states to your present state, you will see continuous ‘advancement’ toward your present condition—”

Obert: “*Wander on a random path?*”

Subhan: “I’m just pointing out that saying, ‘Look how much better things are now’, when your criterion for ‘better’ is comparing past moral values to yours, does not establish any directional trend in human progress.”

Obert: “Your strange beliefs about the nature of morality have destroyed your soul. I don’t even believe in souls, and I’m saying that.”

Subhan: “Look, depending on which arguments do, in fact, move us, you might be able to regard the process of changing terminal values as a directional progress. You might be able to show that the change had a consistent trend as we thought of more and more arguments. But that doesn’t show that morality is something *outside* us. We could even—though this is psychologically unrealistic—choose to *regard you* as computing a converging approximation to your ‘ideal wants’, so that you would have meta-values that defined both your present value and the rules for updating them. But these would be *your* meta-values and *your* ideals and *your* computation, just as much as pepperoni is *your own* taste in pizza toppings. You may not know your *real* favorite ever pizza topping, until you’ve tasted many possible flavors.”

Obert: “Leaving out *what* it is that you just compared to pizza toppings, I begin to be suspicious of the all-embracingness of your viewpoint. No matter *what* my mind does, you can simply call it a still-more-modified ‘want’. I think that *you* are the one suffering from meta-level confusion, not I. Appealing to right is not the same as appealing to desire. Just because the appeal is judged *inside*

*my brain*, doesn't mean that the appeal is not *to* something more than my desires. Why can't my brain compute duties as well as desires?"

Subhan: "What is the difference between duty and desire?"

Obert: "A duty is something you must do whether you want to or not."

Subhan: "Now you're just being incoherent. Your brain computes something it wants to do whether it wants to or not?"

Obert: "No, *you* are the one whose theory makes this incoherent. Which is why your theory ultimately fails to add up to morality."

Subhan: "I say again that you underestimate the power of mere wanting. And more: *You accuse me* of incoherence? You say that *I* suffer from meta-level confusion?"

Obert: "Er... yes?"

*To be continued...*

## 10. Is Morality Given?<sup>1</sup>

**Continuation of:** Is Morality Preference?<sup>2</sup>

(Disclaimer: Neither Subhan nor Obert represent my own position on morality; rather they represent different sides of the *questions* I hope to answer.)

Subhan: “What is this ‘morality’ stuff, if it is *not* a preference within you?”

Obert: “I know that my mere wants, don’t change what is *right*; but I don’t claim to have absolute knowledge of what is right—”

Subhan: “You’re not escaping that easily! How does a universe in which murder is wrong, differ from a universe in which murder is right? How can you detect the difference experimentally? If the answer to that is ‘No’, then how does any human being come to *know* that murder is wrong?”

Obert: “Am I allowed to say ‘I don’t know?’”

Subhan: “No. You believe *now* that murder is wrong. You must believe you *already* have evidence and you should be able to present it *now*.”

Obert: “That’s too strict! It’s like saying to a hunter-gatherer, ‘Why is the sky blue?’ and expecting an immediate answer.”

Subhan: “No, it’s like saying to a hunter-gatherer: Why do you *believe* the sky is blue?”

Obert: “Because it seems blue, just as murder seems wrong. Just don’t ask me what the sky is, or how I can see it.”

Subhan: “But—aren’t we discussing the nature of morality?”

Obert: “That, I confess, is not one of my strong points. I specialize in plain old morality. And as a matter of morality, I know that I can’t make murder *right* just by wanting to kill someone.”

Subhan: “But if you *wanted* to kill someone, you would say, ‘I know murdering this guy is right, and I couldn’t make it wrong just by not wanting to do it.’”

Obert: “Then, if I said that, I would be wrong. That’s common moral sense, right?”

---

1. [http://lesswrong.com/lw/ry/is\\_morality\\_given/](http://lesswrong.com/lw/ry/is_morality_given/)

2. Page 871, ‘Is Morality Preference?’.

Subhan: “Argh! It’s difficult to even argue with you, since you won’t tell me exactly what you think morality is made of, or where you’re getting all these amazing moral truths—”

Obert: “Well, I do regret having to frustrate you. But it’s more important that I *act morally*, than that I come up with amazing new theories of the *nature* of morality. I don’t claim that my strong point is in explaining the fundamental nature of morality. Rather, my strong point is coming up with theories of morality that give normal moral answers to questions like, ‘If you feel like killing someone, does that make it right to do so?’ The common-sense answer is ‘No’ and I really see no reason to adopt a theory that makes the answer ‘Yes’. Adding up to moral normality—that is my theory’s strong point.”

Subhan: “Okay... look. You say that, if you believed it was right to murder someone, you would be *wrong*.”

Obert: “Yes, of course! And just to cut off any quibbles, we’ll specify that we’re not talking about going back in time and shooting Stalin, but rather, stalking some innocent bystander through a dark alley and slitting their throat for no other reason but my own enjoyment. That’s *wrong*.”

Subhan: “And *anyone* who says murder is right, is mistaken.”

Obert: “Yes.”

Subhan: “Suppose there’s an alien species somewhere in the vastness of the multiverse, who evolved from carnivores. In fact, through most of their evolutionary history, they were cannibals. They’ve evolved different emotions from us, and they have no concept that murder is wrong—”

Obert: “Why doesn’t their society fall apart in an orgy of mutual killing?”

Subhan: “That doesn’t matter for our purposes of theoretical metaethical investigation. But since you ask, we’ll suppose that the Space Cannibals have a strong sense of *honor*—they won’t kill someone they promise not to kill; they have a very strong idea that violating an oath is wrong. Their society holds together on that basis, and on the basis of vengeance contracts with private assassination companies. But so far as the actual killing is concerned, the aliens just think it’s fun. When someone gets executed for, say,

driving through a traffic light, there's a bidding war for the rights to personally tear out the offender's throat."

Obert: "Okay... where is this going?"

Subhan: "I'm proposing that the Space Cannibals not only have no sense that murder is wrong—indeed, they have a positive sense that killing is an important part of life—but moreover, there's no path of arguments you could use to *persuade* a Space Cannibal of your view that murder is wrong. There's no fact the aliens can learn, and no chain of reasoning they can discover, which will *ever* cause them to conclude that murder is a moral wrong. Nor is there any way to persuade them that they *should* modify themselves to perceive things differently."

Obert: "I'm not sure I believe *that's* possible—"

Subhan: "Then you believe in universally compelling arguments<sup>3</sup> processed by a ghost in the machine<sup>4</sup>. For every possible mind<sup>5</sup> whose utility function assigns terminal value<sup>6</sup> +1, mind design space<sup>7</sup> contains an equal and opposite mind whose utility function assigns terminal value—1. A mind is a physical device and you can't have a little blue woman pop out of nowhere and make it say 1 when the physics calls for it to say 0."

Obert: "Suppose I were to concede this. Then?"

Subhan: "Then it's possible to have an alien species that believes murder is not wrong, and moreover, will continue to believe this given knowledge of every possible fact and every possible argument. Can you say these aliens are *mistaken*?"

Obert: "Maybe it's the right thing to do in *their* very different, alien world—"

Subhan: "And then they land on Earth and start slitting human throats, laughing all the while, because they don't believe it's wrong. Are they *mistaken*?"

Obert: "Yes."

Subhan: "Where exactly is the mistake? In which step of reasoning?"

3. Page 844, 'No Universally Compelling Arguments'.

4. [http://lesswrong.com/lw/rf/ghosts\\_in\\_the\\_machine/](http://lesswrong.com/lw/rf/ghosts_in_the_machine/)

5. [http://lesswrong.com/lw/rm/the\\_design\\_space\\_of\\_mindsingeneral/](http://lesswrong.com/lw/rm/the_design_space_of_mindsingeneral/)

6. [http://lesswrong.com/lw/l4/terminal\\_values\\_and\\_instrumental\\_values/](http://lesswrong.com/lw/l4/terminal_values_and_instrumental_values/)

7. [http://lesswrong.com/lw/rm/the\\_design\\_space\\_of\\_mindsingeneral/](http://lesswrong.com/lw/rm/the_design_space_of_mindsingeneral/)

Obert: “I don’t know exactly. My guess is that they’ve got a bad axiom.”

Subhan: “Dammit! Okay, look. Is it possible that—by analogy with the Space Cannibals—there are true moral facts of which the human species is not only *presently* unaware, but incapable of perceiving *in principle*? Could we have been born defective—incapable even of being *compelled* by the arguments that would lead us to the light? Moreover, born without any desire to modify ourselves to be capable of understanding such arguments? Could we be *irrevocably mistaken* about morality—just like you say the Space Cannibals are?”

Obert: “I... guess so...”

Subhan: “You guess so? Surely this is an inevitable consequence of believing that morality is a given, independent of anyone’s preferences! Now, is it possible that *we*, not the Space Cannibals, are the ones who are irrevocably mistaken in believing that murder is wrong?”

Obert: “*That* doesn’t seem likely.”

Subhan: “I’m not asking you if it’s likely, I’m asking you if it’s *logically possible!* If it’s *not* possible, then you have just confessed that human morality is ultimately determined by our human constitutions. And if it *is* possible, then what distinguishes this scenario of ‘humanity is irrevocably mistaken about morality’, from finding a stone tablet on which is written the phrase ‘Thou Shalt Murder’ without any known justification attached? How is a given morality any different from an unjustified stone tablet?”

Obert: “Slow down. Why does this argument show that morality is determined by our own constitutions?”

Subhan: “Once upon a time, theologians tried to say that God was the foundation of morality. And even since the time of the ancient Greeks, philosophers were sophisticated enough to go on and ask the next question—‘*Why* follow God’s commands?<sup>8</sup>’ Does God have *knowledge* of morality, so that we should follow Its orders as good advice? But then what is this morality, outside God, of which God has knowledge? Do God’s commands *determine* morality? But then why, *morally*, should one follow God’s orders?”

---

8. <http://plato.stanford.edu/entries/metaethics/#EutPro>

Obert: "Yes, this demolishes attempts to answer questions about the nature of morality just by saying 'God!', unless you answer the obvious further questions. But so what?"

Subhan: "And furthermore, let us castigate those who made the argument originally, for the sin of trying to *cast off responsibility*—trying to wave a scripture and say, 'I'm just following God's orders!' Even if God *had* told them to do a thing, it would still have been *their own decision* to follow God's orders."

Obert: "I agree—as a matter of morality, there is no evading of moral responsibility. Even if your parents, or your government, or some kind of hypothetical superintelligence, tells you to do something, you are responsible for your decision<sup>9</sup> in doing it."

Subhan: "But you see, this also demolishes the idea of any morality that is outside, beyond, or above human preference. Just substitute 'morality' for 'God' in the argument!"

Obert: "What?"

Subhan: "John McCarthy<sup>10</sup> said: 'You say you couldn't live if you thought the world had no purpose. You're saying that you can't form purposes of your own—that you need someone to tell you what to do. The average child has more gumption than that.' For every kind of stone tablet that you might imagine anywhere, in the trends of the universe or in the structure of logic, you are still left with the question: 'And *why* obey this morality?' It would be *your decision* to follow this trend of the universe, or obey this structure of logic. Your decision—and *your preference*."

Obert: "That doesn't follow! Just because it is *my decision* to be moral—and even because there are drives in me that lead me to make that decision—it doesn't follow that the morality I follow consists *merely* of my preferences. If someone gives me a pill that makes me prefer to *not* be moral, to commit murder, then this just alters my preference—but *not* the morality; murder is still wrong. That's common moral sense—"

Subhan: "I beat my head against my keyboard! What about *scientific* common sense? If morality is this mysterious *given* thing, from beyond space and time—and I don't even see why we *should* follow it, in that case—but in any case, if morality exists indepen-

9. Page 321, 'Policy Debates Should Not Appear One-Sided'.

10. [http://en.wikiquote.org/wiki/John\\_McCarthy](http://en.wikiquote.org/wiki/John_McCarthy)

dently of human nature, then isn't it a *remarkable coincidence* that, say, *love is good?*"

Obert: "Coincidence? How so?"

Subhan: "Just where on Earth do you think the emotion of *love* comes from? If the ancient Greeks had ever thought of the theory of natural selection, they could have looked at the human institution of sexual romance, or parental love for that matter, and deduced in one flash that human beings had evolved—or at least derived tremendous Bayesian evidence for human evolution. Parental bonds and sexual romance clearly display the signature of evolutionary psychology<sup>11</sup>—they're archetypal cases, in fact, so obvious we usually don't even see it."

Obert: "But love isn't just about reproduction—"

Subhan: "Of course not; individual organisms are adaptation-executors, not fitness-maximizers<sup>12</sup>. But for something independent of humans, morality looks remarkably like godshatter of natural selection<sup>13</sup>. Indeed, it is far too much coincidence for me to credit. Is happiness morally preferable to pain? What a coincidence! And if you claim that there is any emotion, any instinctive preference, any complex brain circuitry in humanity which was created by some external morality thingy and not natural selection, then you are infringing upon science and you will surely be torn to shreds—science has never needed to postulate anything but evolution to explain any feature of human psychology—"

Obert: "I'm *not* saying that humans got here by anything except evolution."

Subhan: "Then why does morality look so amazingly like a product of an evolved psychology?"

Obert: "I don't claim perfect access to moral truth; maybe, being human, I've made certain mistakes about morality—"

Subhan: "Say *that*—forsake love and life and happiness, and follow some useless damn trend of the universe or whatever—and you will lose every scrap of the moral normality that you once touted as your strong point. And I will be right here, asking, 'Why even bother?' It would be a pitiful mind indeed that demanded author-

11. [http://lesswrong.com/lw/l1/evolutionary\\_psychology/](http://lesswrong.com/lw/l1/evolutionary_psychology/)

12. [http://lesswrong.com/lw/lo/adaptationexecutors\\_not\\_fitnessmaximizers/](http://lesswrong.com/lw/lo/adaptationexecutors_not_fitnessmaximizers/)

13. [http://lesswrong.com/lw/l3/thou\\_art\\_godshatter/](http://lesswrong.com/lw/l3/thou_art_godshatter/)

itative answers so strongly, that it would forsake all good things to have some authority beyond itself to follow.”

Obert: “All right... then maybe the reason morality seems to bear certain similarities to our human constitutions, is that we could only perceive morality at all, if we happened, by luck, to evolve in consonance with it.”

Subhan: “Horsemanure.”

Obert: “Fine... you’re right, that wasn’t very plausible. Look, I admit you’ve driven me into quite a corner here. But even if there *were* nothing more to morality than preference, I would still prefer to act as morality were real. I mean, if it’s all just preference, that way is as good as anything else—”

Subhan: “Now you’re just trying to avoid facing reality<sup>14</sup>! Like someone who says, ‘If there is no Heaven or Hell, then I may as well still act as if God’s going to punish me for sinning.’”

Obert: “That may be a good metaphor, in fact. Consider two theists, in the process of becoming atheists. One says, ‘There is no Heaven or Hell, so I may as well cheat and steal, if I can get away without being caught, since there’s no God to watch me.’ And the other says, ‘Even though there’s no God, I intend to *pretend* that God is watching me, so that I can go on being a moral person.’ Now they are both mistaken, but the first is straying much further from the path.”

Subhan: “And what is the second one’s flaw? *Failure to accept personal responsibility!*”

Obert: “Well, and I admit I find that a more compelling argument than anything else you have said. Probably because it is a moral argument, and it has always been morality, not metaethics, with which I claimed to be concerned. But even so, after our whole conversation, I still maintain that wanting to murder someone does not make murder *right*. Everything that you have said about preference is interesting, but it is ultimately *about* preference—about minds and what they are designed to desire—and not about this other thing that humans sometimes talk about, ‘morality’. I can just ask Moore’s Open Question<sup>15</sup>: Why should I *care* about human preferences? What makes following human preferences *right*? By

<sup>14</sup>. Page 639, ‘You Can Face Reality’.

<sup>15</sup>. <http://plato.stanford.edu/entries/metaethics/#IsOOpeQueArg>

changing a mind, you can change what it prefers; you can even change what it *believes* to be right; but you cannot change what *is* right. Anything you talk about, that can be changed in this way, is not ‘right-ness’.”

Subhan: “So you take refuge in arguing from definitions<sup>16</sup>? ”

Obert: “You know, when I reflect on this whole argument, it seems to me that your position has the definite advantage when it comes to arguments about ontology and reality and all that stuff—”

Subhan: “*All that stuff?* What else *is* there, besides reality?”

Obert: “Okay, the morality-as-preference viewpoint is a lot easier to shoehorn into a universe of quarks. But I still think the morality-as-given viewpoint has the advantage when it comes to, you know, the actual *morality* part of it—giving answers that are good in the sense of being *morally* good, not in the sense of being a good reductionist. Because, you know, there *are* such things as moral errors, there *is* moral progress, and you really *shouldn’t* go around thinking that murder would be right if you wanted it to be right.”

Subhan: “That sounds to me like the logical fallacy of appealing to consequences.”

Obert: “Oh? Well, it sounds to *me* like an incomplete reduction—one that doesn’t quite add up to normality.”

---

<sup>16</sup> Page 263, ‘Arguing “By Definition”’.

## 11. Where Recursive Justification Hits Bottom<sup>1</sup>

**Followup to:** No Universally Compelling Arguments<sup>2</sup>, Passing the Recursive Buck<sup>3</sup>, Wrong Questions<sup>4</sup>, A Priori<sup>5</sup>

Why do I believe that the Sun will rise tomorrow?

Because I've seen the Sun rise on thousands of previous days.

Ah... but why do I believe the future will be like the past?

Even if I go past the mere surface observation of the Sun rising, to the apparently universal and exceptionless<sup>6</sup> laws of gravitation and nuclear physics, then I am still left with the question: "Why do I believe this will also be true tomorrow?"

I could appeal to Occam's Razor<sup>7</sup>, the principle of using the simplest theory that fits the facts... but why believe in Occam's Razor? Because it's been successful on past problems? But who says that this means Occam's Razor will work tomorrow?

And lo, the one said:

"Science also depends on unjustified assumptions. Thus science is ultimately based on faith, *so don't you criticize me for believing in [silly-belief-#238721].*"

As I've previously observed<sup>8</sup>:

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-triumphant tone, rather than as a compliment?

---

1. [http://lesswrong.com/lw/so/where\\_recursive\\_justification\\_hits\\_bottom/](http://lesswrong.com/lw/so/where_recursive_justification_hits_bottom/)

2. Page 844, 'No Universally Compelling Arguments'.

3. [http://lesswrong.com/lw/rd/passing\\_the\\_recursive\\_buck/](http://lesswrong.com/lw/rd/passing_the_recursive_buck/)

4. Page 679, 'Wrong Questions'.

5. [http://lesswrong.com/lw/k2/a\\_priori/](http://lesswrong.com/lw/k2/a_priori/)

6. [http://lesswrong.com/lw/br/universal\\_law/](http://lesswrong.com/lw/br/universal_law/)

7. Page 95, 'Occam's Razor'.

8. Page 599, 'The Fallacy of Gray'.

Arguing that you should be immune to criticism is rarely a good sign.

But this doesn't answer the legitimate philosophical dilemma<sup>9</sup>: If every belief must be justified, and those justifications in turn must be justified, then how is the infinite recursion terminated?

And if you're allowed to end in something assumed-without-justification, then why aren't you allowed to assume *anything* without justification?

A similar critique is sometimes leveled against Bayesianism—that it requires assuming some prior—by people who apparently think that the problem of induction is a *particular* problem of Bayesianism, which you can avoid by using classical statistics. I will speak of this later, perhaps.

But first, let it be clearly admitted that the rules of Bayesian updating, do *not* of themselves solve the problem of induction.

Suppose you're drawing red and white balls from an urn<sup>10</sup>. You observe that, of the first 9 balls, 3 are red and 6 are white. What is the probability that the next ball drawn will be red?

That depends on your prior beliefs about the urn. If you think the urn-maker generated a uniform random number between 0 and 1, and used that number as the fixed probability of each ball being red, then the answer is 4/11 (by Laplace's Law of Succession). If you think the urn originally contained 10 red balls and 10 white balls, then the answer is 7/11.

Which goes to say that, with the right prior—or rather the wrong prior—the chance of the Sun rising tomorrow, would seem to go *down* with each succeeding day... if you were absolutely certain, *a priori*, that there was a great barrel out there from which, on each day, there was drawn a little slip of paper that determined whether the Sun rose or not; and that the barrel contained only a limited number of slips saying “Yes”, and the slips were drawn without replacement.

There are possible minds in mind design space<sup>11</sup> who have anti-Occamian and anti-Laplacian priors; they believe that simpler theo-

---

9. Page 679, ‘Wrong Questions’.

10. [http://lesswrong.com/lw/bk/priors\\_as\\_mathematical\\_objects/](http://lesswrong.com/lw/bk/priors_as_mathematical_objects/)

11. [http://lesswrong.com/lw/rm/the\\_design\\_space\\_of\\_mindsgeneral/](http://lesswrong.com/lw/rm/the_design_space_of_mindsgeneral/)

ries are less likely to be correct, and that the more often something happens, the less likely it is to happen again.

And when you ask these strange beings why they keep using priors that never seem to work in real life... they reply, “Because it’s never worked for us before!”

Now, one lesson you might derive from this, is “Don’t be born with a stupid prior.” This is an amazingly helpful principle on many real-world problems, but I doubt it will satisfy philosophers.

Here’s how I treat this problem myself: I try to approach questions like “Should I trust my brain?” or “Should I trust Occam’s Razor?” as though they were *nothing special*—or at least, nothing special as deep questions go.

Should I trust Occam’s Razor? Well, how well does (any particular version of) Occam’s Razor seem to work in practice? What kind of probability-theoretic justifications<sup>12</sup> can I find for it? When I look at the universe, does it seem like the kind of universe in which Occam’s Razor would work well?

Should I trust my brain? Obviously not; it doesn’t always work. But nonetheless, the human brain seems much more powerful than the most sophisticated computer programs I could consider trusting otherwise. How well does my brain work in practice, on which sorts of problems?

When I examine the causal history of my brain—its origins<sup>13</sup> in natural selection<sup>14</sup>—I find, on the one hand, all sorts of specific reasons for doubt; my brain was optimized to run on the ancestral savanna, not to do math. But on the other hand, it’s also clear why, loosely speaking, it’s possible that the brain really could work. Natural selection would have quickly eliminated brains so *completely* unsuited to reasoning, so *anti-helpful*, as anti-Occamian or anti-Laplacian priors.

So what I did in practice, does *not* amount to declaring a sudden halt<sup>15</sup> to questioning and justification. I’m not halting the chain of examination at the point that I encounter Occam’s Razor, or my brain, or some other unquestionable. The chain of examination continues—but it continues, unavoidably, using my current brain

12. [http://lesswrong.com/lw/q3/decoherence\\_is\\_simple/](http://lesswrong.com/lw/q3/decoherence_is_simple/)

13. [http://lesswrong.com/lw/l1/evolutionary\\_psychology/](http://lesswrong.com/lw/l1/evolutionary_psychology/)

14. [http://lesswrong.com/lw/kr/an\\_alien\\_god/](http://lesswrong.com/lw/kr/an_alien_god/)

15. Page 175, ‘Explain/Worship/Ignore?’.

and my current grasp on reasoning techniques. *What else could I possibly use?*

Indeed, no matter *what* I did with this dilemma, it would be me doing it. Even if I trusted something else, like some computer program, it would be my own decision to trust it.

The technique of rejecting beliefs that have absolutely no justification, is in general an extremely important one. I sometimes say that the fundamental question of rationality is “Why do you believe what you believe?” I don’t even want to say something that *sounds* like it might allow a single exception to the rule that everything needs justification.

Which is, itself, a dangerous sort of motivation; you can’t always avoid everything that might be risky, and when someone annoys you by saying something silly, you can’t reverse that stupidity to arrive at intelligence<sup>16</sup>.

But I would nonetheless emphasize the difference between saying:

“Here is this assumption I cannot justify, which must be simply taken, and not further examined.”

Versus saying:

“Here the inquiry continues to examine this assumption, with the full force of my *present intelligence*—as opposed to the full force of something else, like a random number generator or a magic 8-ball—even though my present intelligence happens to be founded on this assumption.”

Still... wouldn’t it be nice if we could examine the problem of how much to trust our brains *without* using our current intelligence? Wouldn’t it be nice if we could examine the problem of how to think, *without* using our current grasp of rationality?

When you phrase it *that* way, it starts looking like the answer might be “No”.

E. T. Jaynes used to say that you must always use all the information available to you—he was a Bayesian probability theorist, and

---

<sup>16</sup> Page 335, ‘Reversed Stupidity Is Not Intelligence’.

had to clean up the paradoxes other people generated when they used different information at different points in their calculations. The principle of “*Always put forth your true best effort*” has at least as much appeal as “*Never do anything that might look circular.*“ After all, the alternative to putting forth your best effort is presumably doing less than your best.

*But still...* wouldn’t it be nice if there were some way to justify using Occam’s Razor, or justify predicting that the future will resemble the past, *without* assuming that those methods of reasoning which have worked on previous occasions are better than those which have continually failed?

Wouldn’t it be nice if there were some chain of justifications that neither ended in an unexamining assumption, nor was forced to examine itself under its own rules, but, instead, could be explained starting from absolute scratch to an ideal philosophy student of perfect emptiness<sup>17</sup>?

Well, I’d certainly be interested, but I don’t expect to see it done any time soon. I’ve argued elsewhere in several places against the idea that you can have a perfectly empty ghost-in-the-machine<sup>18</sup>; there is no argument that you can explain to a rock<sup>19</sup>.

Even if someone cracks the First Cause problem<sup>20</sup> and comes up with *the actual reason the universe is simple, which does not itself presume a simple universe...* then I would still expect that the explanation could only be understood by a mindful listener, and not by, say, a rock. A listener that didn’t start out already implementing modus ponens<sup>21</sup> might be out of luck.

So, at the end of the day, what happens when someone keeps asking me “Why do you believe what you believe?”

At present, I start going around in a loop at the point where I explain, “I predict the future as though it will resemble the past on the simplest and most stable level of organization I can identify, because previously, this rule has usually worked to generate good results; and using the simple assumption of a simple universe, I can see *why* it generates good results; and I can even see how my brain

17. Page 844, ‘No Universally Compelling Arguments’.

18. Page 844, ‘No Universally Compelling Arguments’.

19. Page 861, ‘Created Already In Motion’.

20. Page 679, ‘Wrong Questions’.

21. Page 861, ‘Created Already In Motion’.

might have evolved to be able to observe the universe with some degree of accuracy, if my observations are correct.”

But then... haven’t I just licensed *circular logic*?

Actually, I’ve just licensed *reflecting on your mind’s degree of trustworthiness, using your current mind as opposed to something else*.

Reflection of this sort is, indeed, the reason we reject most circular logic in the first place. We want to have a coherent causal story about how our mind comes to know something<sup>22</sup>, a story that explains how the process we used to arrive at our beliefs, is itself trustworthy. This is the essential demand behind the rationalist’s fundamental question, “Why do you believe what you believe?”

Now suppose you write on a sheet of paper: “(1) Everything on this sheet of paper is true, (2) The mass of a helium atom is 20 grams.” If that trick actually *worked in real life*, you would be able to know the true mass of a helium atom just by believing some circular logic which asserted it. Which would enable you to arrive at a true map of the universe sitting in your living room with the blinds drawn. Which would violate the second law of thermodynamics<sup>23</sup> by generating information from nowhere. Which would not be a plausible story about how your mind could end up believing something true.

*Even if* you started out believing the sheet of paper, it would not seem that you had any reason for why the paper corresponded to reality. It would just be a miraculous coincidence<sup>24</sup> that (a) the mass of a helium atom was 20 grams, and (b) the paper happened to say so.

Believing, in general, self-validating statement sets, does not seem like it should work to map external reality—when we *reflect on it as a causal story about minds*—using, of course, our *current* minds to do so.

But what about evolving to give more credence to simpler beliefs, and to believe that algorithms which have worked in the past are more likely to work in the future? *Even when* we reflect on this as a causal story of the origin of minds, it still seems like this could plausibly work to map reality.

22. Page 99, ‘The Lens That Sees Its Flaws’.

23. [http://lesswrong.com/lw/05/the\\_second\\_law\\_of\\_thermodynamics\\_and\\_engines\\_of/](http://lesswrong.com/lw/05/the_second_law_of_thermodynamics_and_engines_of/)

24. [http://lesswrong.com/lw/06/perpetual\\_motion\\_beliefs/](http://lesswrong.com/lw/06/perpetual_motion_beliefs/)

And what about trusting reflective coherence in general? Wouldn't most possible minds, randomly generated and allowed to settle into a state of reflective coherence, be incorrect? Ah, but *we* evolved by natural selection; we were not generated randomly.

If trusting this argument seems worrisome to you, then forget about the problem of philosophical justifications, and ask yourself whether it's really truly true.

(You will, of course, use your own mind to do so.)

Is this the same as the one who says, "I believe that the Bible is the word of God, because the Bible says so"?

Couldn't they argue that their blind faith must also have been placed in them by God, and is therefore trustworthy?

In point of fact, when religious people finally come to reject the Bible, they do *not* do so by magically jumping to a non-religious state of pure emptiness, and then evaluating their religious beliefs in that non-religious state of mind, and then jumping back to a new state with their religious beliefs removed.

People go from being religious, to being non-religious, because even in a religious state of mind, doubt seeps in. They notice their prayers (and worse, the prayers of seemingly much worthier people) are not being answered. They notice that God, who speaks to them in their heart in order to provide seemingly consoling answers about the universe, is not able to tell them the hundredth digit of pi (which would be a lot more reassuring, if God's purpose were reassurance). They examine the story of God's creation of the world and damnation of unbelievers, and it doesn't seem to make sense even under their own religious premises.

Being religious doesn't make you less than human. Your brain still has the abilities of a human brain. The dangerous part is that being religious might stop you from *applying* those native abilities to your religion—stop you from *reflecting fully* on yourself. People don't heal their errors by resetting themselves to an ideal philosopher of pure emptiness and reconsidering all their sensory experiences from scratch. They heal themselves by becoming more willing to question their current beliefs, using more of the power of their current mind.

This is why it's important to distinguish between *reflecting on your mind using your mind* (it's not like you can use anything else) and *having an unquestionable assumption that you can't reflect on*.

"I believe that the Bible is the word of God, because the Bible says so." Well, if the Bible *were* an astoundingly reliable source of information about all other matters, if it had not said that grasshoppers had four legs or that the universe was created in six days, but had instead contained the Periodic Table of Elements centuries before chemistry—if the Bible had served us only well and told us only truth—then we might, in fact, be inclined to take seriously the additional statement in the Bible, that the Bible had been generated by God. We might not trust it entirely, because it could also be aliens or the Dark Lords of the Matrix, but it would at least be worth taking seriously.

Likewise, if everything *else* that priests had told us, turned out to be true, we might take more seriously their statement that faith had been placed in us by God and was a systematically trustworthy source—especially if people could divine the hundredth digit of pi by faith as well.

So the important part of appreciating the circularity of "I believe that the Bible is the word of God, because the Bible says so," is not so much that you are going to reject the idea of reflecting on your mind using your current mind. But, rather, that you realize that anything which calls into question the Bible's trustworthiness, also calls into question the Bible's assurance of its trustworthiness.

This applies to rationality too: if the future should cease to resemble the past—even on its lowest and simplest and most stable observed levels of organization—well, mostly, I'd be dead, because my brain's processes require a lawful universe where chemistry goes on working. But if somehow I survived, then I would have to start questioning the principle that the future should be predicted to be like the past.

But for now... what's the *alternative* to saying, "I'm going to believe that the future will be like the past on the most stable level of organization I can identify, because that's previously worked better for me than any other algorithm I've tried"?

Is it saying, "I'm going to believe that the future will *not* be like the past, because that algorithm has always failed before"?

At this point I feel obliged to drag up the point that rationalists are not out to win arguments with ideal philosophers of perfect emptiness; we are simply out to win<sup>25</sup>. For which purpose we want to get as close to the truth as we can possibly manage. So at the end of the day, I embrace the principle: “Question your brain, question your intuitions, question your principles of rationality, *using the full current force of your mind, and doing the best you can do at every point.*”

If one of your current principles does come up wanting—according to your own mind’s examination, since you can’t step outside yourself<sup>26</sup>—then change it! And then go back and look at things again, using your new improved principles.

The point is not to be reflectively consistent. The point is to win. But if you look at yourself and play to win, you are making yourself more reflectively consistent—that’s what it means to “play to win” while “looking at yourself”.

Everything, without exception, needs justification. Sometimes—unavoidably, as far as I can tell—those justifications will go around in reflective loops. I do think that reflective loops have a meta-character which should enable one to distinguish them, by common sense, from circular logics. But anyone seriously considering a circular logic in the first place, is probably out to lunch in matters of rationality; and will simply insist that their circular logic is a “reflective loop” even if it consists of a single scrap of paper saying “Trust me”. Well, you can’t always optimize your rationality techniques according to the sole consideration of preventing those bent on self-destruction from abusing them.

The important thing is to *hold nothing back* in your criticisms of how to criticize; nor should you regard the unavoidability of loopy justifications as a warrant of *immunity from questioning*.

Always apply full force, whether it loops or not—do the best you can possibly do, whether it loops or not—and play, ultimately, to win.

---

<sup>25</sup>. Page 644, ‘Something to Protect’.

<sup>26</sup>. [http://lesswrong.com/lw/r/the\\_ultimate\\_source/](http://lesswrong.com/lw/r/the_ultimate_source/)

## 12. My Kind of Reflection<sup>1</sup>

### Followup to: Where Recursive Justification Hits Bottom<sup>2</sup>

In “Where Recursive Justification Hits Bottom<sup>3</sup>”, I concluded that it’s okay to use induction to reason about the probability that induction will work in the future, given that it’s worked in the past; or to use Occam’s Razor to conclude that the simplest explanation for why Occam’s Razor works is that the universe itself is fundamentally simple.

Now I am far from the first person to consider reflective application of reasoning principles. Chris Hibbert compared my view to Bartley’s Pan-Critical Rationalism (I was wondering whether that would happen). So it seems worthwhile to state what I see as the distinguishing features of my view of reflection, which may or may not happen to be shared by any other philosopher’s view of reflection.

- All of my philosophy here *actually* comes from trying to figure out how to build a self-modifying AI that applies its own reasoning principles to itself in the process of rewriting its own source code. So whenever I talk about using induction to license induction, I’m *really* thinking about an inductive AI considering a rewrite of the part of itself that performs induction. If you wouldn’t want the AI to rewrite its source code to not use induction, your philosophy had better not label induction as unjustifiable.

- One of the most powerful general principles I know for AI in general, is that the true Way generally turns out to be *naturalistic*—which for reflective reasoning, means treating transistors inside the AI, just as if they were transistors found in the environment; *not* an ad-hoc special case. This is the real source of my insistence in “Recursive Justification” that questions like “How well does my version of Occam’s Razor work?” should be considered just like an ordinary question—or at least an ordinary very deep question. I strongly suspect that a correctly built AI, in pondering modifications to the part of its source code that implements Occamian reasoning, will not have to do anything special as it ponders—in

---

1. [http://lesswrong.com/lw/s2/my\\_kind\\_of\\_reflection/](http://lesswrong.com/lw/s2/my_kind_of_reflection/)

2. Page 887, ‘Where Recursive Justification Hits Bottom’.

3. Page 887, ‘Where Recursive Justification Hits Bottom’.

particular, it shouldn't have to make a special effort to avoid using Occamian reasoning.

- I don't think that "reflective coherence" or "reflective consistency" should be considered as a desideratum in itself. As I said in the *Twelve Virtues* and the *Simple Truth*, if you make five accurate maps of the same city, then the maps will necessarily be consistent with each other; but if you draw one map by fantasy and then make four copies, the five will be consistent but not accurate. In the same way, no one is deliberately pursuing reflective consistency, and reflective consistency is not a special warrant of trustworthiness; the goal is to win<sup>4</sup>. But anyone who pursues the goal of winning, using their current notion of winning, and modifying their own source code, will end up reflectively consistent as a side effect—just like someone continually striving to improve their map of the world should find the parts becoming more consistent among themselves, as a side effect. If you put on your AI goggles, then the AI, rewriting its own source code, is not trying to make itself "reflectively consistent"—it is trying to optimize the expected utility of its source code, and it happens to be doing this using its current mind's anticipation of the consequences.

- One of the ways I license using induction and Occam's Razor to consider "induction" and "Occam's Razor", is by appealing to E. T. Jaynes's principle that we should always use all the information available to us (computing power permitting) in a calculation. If you think induction works, then you should use it in order to use your maximum power, including when you're thinking about induction.

- In general, I think it's valuable to distinguish a defensive posture where you're imagining how to justify your philosophy to a philosopher that questions you, from an aggressive posture where you're trying to get as close to the truth as possible. So it's not that being suspicious of Occam's Razor, but using your current mind and intelligence to inspect it, shows that you're being *fair* and *defensible* by questioning your foundational beliefs. Rather, the reason why you would inspect Occam's Razor is to see if you could improve your application of it, or if you're worried it might really be wrong. I tend to deprecate mere dutiful<sup>5</sup> doubts<sup>6</sup>.

---

4. [http://lesswrong.com/lw/nc/newcombs\\_problem\\_and\\_regret\\_of\\_rationality/](http://lesswrong.com/lw/nc/newcombs_problem_and_regret_of_rationality/)

5. Page 640, 'The Meditation on Curiosity'.

- If you run around inspecting your foundations, I expect you to actually improve them, not just dutifully investigate. Our brains are built to assess “simplicity” in a certain intuitive way that makes Thor sound simpler than Maxwell’s Equations<sup>7</sup> as an explanation for lightning. But, having gotten a better look at the way the universe really works, we’ve concluded that differential equations (which few humans master) are actually *simpler* (in an information-theoretic sense) than heroic mythology (which is how most tribes explain the universe). This being the case, we’ve tried to import our notions of Occam’s Razor into math as well.

- On the other hand, the improved foundations should still add up to normality<sup>8</sup>;  $2 + 2$  should still end up equalling 4, not something new and amazing and exciting like “fish”.

- I think it’s very important to distinguish between the questions “Why does induction work?” and “Does induction work?” The reason *why the universe itself is regular* is still a mysterious question unto us, for now. Strange speculations here may be temporarily needful. But on the other hand, if you start claiming that the universe *isn’t actually regular*, that the answer to “Does induction work?” is “No!”, then you’re wandering into  $2 + 2 = 3$  territory. You’re trying too hard to make your philosophy interesting, instead of correct. An inductive AI asking what probability assignment to make on the next round is asking “*Does* induction work?”, and this is the question that it may answer by inductive reasoning. If you ask “*Why* does induction work?” then answering “Because induction works” is circular logic, and answering “Because I believe induction works” is magical thinking.

- I don’t think that going around in a loop of justifications through the meta-level is the same thing as circular logic. I think the notion of “circular logic” applies within the object level, and is something that is definitely bad and forbidden, on the object level. Forbidding *reflective coherence* doesn’t sound like a good idea. But I haven’t yet sat down and formalized the exact difference—my reflective theory is something I’m trying to work out, not something I have in hand.

---

6. Page 636, ‘The Proper Use of Doubt’.

7. Page 95, ‘Occam’s Razor’.

8. [http://lesswrong.com/lw/qz/living\\_in\\_many\\_worlds/](http://lesswrong.com/lw/qz/living_in_many_worlds/)

## 13. The Genetic Fallacy<sup>1</sup>

In lists<sup>2</sup> of<sup>3</sup> logical<sup>4</sup> fallacies<sup>5</sup>, 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<sup>6</sup> 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<sup>7</sup>. Genetic accusations have a force among humans that they would not have among ideal Bayesians.

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

1. [http://lesswrong.com/lw/s3/the\\_genetic\\_fallacy/](http://lesswrong.com/lw/s3/the_genetic_fallacy/)

2. [http://www.nizkor.org/features/fallacies/genetic\\_fallacy.html](http://www.nizkor.org/features/fallacies/genetic_fallacy.html)

3. [http://en.wikipedia.org/wiki/List\\_of\\_fallacies](http://en.wikipedia.org/wiki/List_of_fallacies)

4. [http://atheism.about.com/library/FAQs/skepticism/blfaq\\_fall\\_genetic.htm](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 469, ‘We Change Our Minds Less Often Than We Think’.

level) the idea that the Bible was written by the hand of God; and who nonetheless think that the Bible contains indispensable ethical wisdom<sup>8</sup>. 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<sup>9</sup>; 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<sup>10</sup> cached<sup>11</sup> 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-cut evidence is what Science<sup>12</sup> 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<sup>13</sup> 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.

8. [http://lesswrong.com/lw/i8/religions\\_claim\\_to\\_be\\_nondisprovable/](http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/)

9. Page 335, ‘Reversed Stupidity Is Not Intelligence’.

10. Page 511, ‘One Argument Against An Army’.

11. Page 451, ‘Cached Thoughts’.

12. [http://lesswrong.com/lw/in/scientific\\_evidence\\_legal\\_evidence\\_rational/](http://lesswrong.com/lw/in/scientific_evidence_legal_evidence_rational/)

13. [http://lesswrong.com/lw/jo/einsteins\\_arrogance/](http://lesswrong.com/lw/jo/einsteins_arrogance/)

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...
- Be extremely suspicious if you find that you still believe the early suggestions of a source you later rejected.

**Added:** Hal Finney suggests<sup>14</sup> that we should call it “the genetic heuristic”.

---

<sup>14</sup>. <http://www.overcomingbias.com/2008/07/genetic-fallacy.html#comment-121978890>

## 14. Fundamental Doubts<sup>1</sup>

**Followup to:** The Genetic Fallacy<sup>2</sup>, Where Recursive Justification Hits Bottom<sup>3</sup>

Yesterday I said that—because humans are not perfect Bayesians—the genetic fallacy is not entirely a fallacy; when new suspicion is cast on one of your fundamental sources, you really *should* doubt all the branches and leaves of that root, even if they *seem* to have accumulated new evidence in the meanwhile.

This is one of the most difficult techniques of rationality (on which I will separately post, one of these days). Descartes, setting out to “doubt, insofar as possible, all things”, ended up trying to prove the existence of God—which, if he wasn’t a secret atheist trying to avoid getting burned at the stake, is pretty pathetic. It is *hard* to doubt an idea to which we are deeply attached; our mind naturally reaches for cached thoughts<sup>4</sup> and rehearsed arguments<sup>5</sup>.

But today’s post concerns a different kind of difficulty—the case where the doubt is so deep, of a source so fundamental, that you *can’t* make a true fresh beginning.

Case in point: Remember when, in the *The Matrix*, Morpheus told Neo that the machines were harvesting the body heat of humans for energy, and liquefying the dead to feed to babies? I suppose you thought something like, “Hey! That violates the second law of thermodynamics.”

Well, it *does* violate the second law of thermodynamics. But if the *Matrix*’s makers had cared about the flaw once it was pointed out to them, they could have fixed the plot hole in any of the sequels, in fifteen seconds, this easily:

Neo: “Doesn’t harvesting human body heat for energy, violate the laws of thermodynamics?”

Morpheus: “Where’d you learn about thermodynamics, Neo?”

---

1. [http://lesswrong.com/lw/s4/fundamental\\_doubts/](http://lesswrong.com/lw/s4/fundamental_doubts/)

2. Page 485, ‘The Genetic Fallacy’.

3. Page 887, ‘Where Recursive Justification Hits Bottom’.

4. Page 451, ‘Cached Thoughts’.

5. Page 511, ‘One Argument Against An Army’.

Neo: "In school."

Morpheus: "Where'd you go to school, Neo?"

Neo: "Oh."

Morpheus: "The machines tell elegant lies."

Now, mind you, I am not saying that this excuses the original mistake in the script. When my mind generated this excuse, it came clearly labeled with that warning sign of which I have spoken<sup>6</sup>, "Tada! Your mind can generate an excuse for *anything*!" You do not need to tell me that my plot-hole-patch is a nitwit idea, I am well aware of that...

...but, in point of fact, if you woke up out of a virtual reality pod one day, you *would* have to suspect all the physics you knew. Even if you looked down and saw that you had hands, you couldn't rely on there being blood and bone inside them. Even if you looked up and saw stars, you couldn't rely on their being trillions of miles away. And even if you found yourself thinking, you couldn't rely on your head containing a brain.

You could still try to doubt, even so. You could do your best to unwind your thoughts past every lesson in school, every science paper read, every sensory experience, every math proof whose seeming approval by other mathematicians might have been choreographed to conceal a subtle flaw...

But suppose you discovered that you were a computer program and that the Dark Lords of the Matrix were actively tampering with your thoughts.

Well... in that scenario, you're pretty much screwed, I'd have to say.

Descartes vastly underestimated the powers of an infinitely powerful deceiving demon when he supposed he could trust "I think therefore I am." Maybe that's just what *they* want you to think. Maybe *they* just inserted that conclusion into your mind with a memory of it seeming to have an irrefutable chain of logical sup-

---

6. [http://lesswrong.com/lw/r3/against\\_devils\\_advocacy/](http://lesswrong.com/lw/r3/against_devils_advocacy/)

port, along with some peer pressure<sup>7</sup> to label it “unquestionable” just like all your friends.

(Personally, I don’t trust “I think therefore I am” even in real life, since it contains a term “am” whose meaning I find confusing, and I’ve learned to spread my confidence intervals very widely in the presence of basic confusion. As for absolute certainty<sup>8</sup>, don’t be silly.)

Every memory of justification could be faked. Every feeling of support could be artificially induced. Modus ponens<sup>9</sup> could be a lie. Your concept of “rational justification”—not just your specific concept, but your notion that any such thing exists at all—could have been manufactured to mislead you. Your trust in Reason itself could have been inculcated to throw you off the trail.

So you might as well not think about the possibility that you’re a brain with choreographed thoughts, because there’s nothing you can do about it...

Unless, of course, that’s what *they* want you to think.

Past a certain level of doubt, it’s not possible to start over fresh. There’s nothing you can *unassume* to find some firm rock on which to stand. You cannot unwind yourself into a perfectly empty and perfectly reliable ghost in the machine<sup>10</sup>.

This level of meta-suspicion should be a rare occasion. For example, suspecting that all academic science is an organized conspiracy<sup>11</sup>, should not run into anything like these meta-difficulties. Certainly, someone does not get to plead that unwinding past the Bible is impossible because it is too foundational; atheists walk the Earth without falling into comas. Remember, when Descartes tried to outwit an infinitely powerful deceiving demon, he first tried to make himself absolutely certain of a highly confusing statement, and then proved the existence of God. Consider that a caution about what you try to claim is “too basic for a fresh beginning”. And even basic things can still be doubted, it is only that we use our untrustworthy brains to doubt them<sup>12</sup>.

7. Page 424, ‘Asch’s Conformity Experiment’.

8. Page 614, ‘Infinite Certainty’.

9. Page 861, ‘Created Already In Motion’.

10. Page 844, ‘No Universally Compelling Arguments’.

11. <http://www.csics.umich.edu/~crshalizi/story-so-far/>

12. Page 887, ‘Where Recursive Justification Hits Bottom’.

Or consider the case of our existence as evolved brains. Natural selection<sup>13</sup> isn't trustworthy, and we have specific reason to suspect it. We know that evolution is stupid<sup>14</sup>. We know many specific ways in which our human brains fail, taken beyond the savanna. But you *can't* clear your mind of evolutionary influences and start over. It would be like deciding that you don't trust neurons, so you're going to clear your mind of brains.

And evolution certainly gets a chance to influence every single thought that runs through your mind! It is the very reason why you exist as a thinker, rather than a lump of carbon—and that doesn't mean evolution summoned a ghost-in-the-machine into you; it *designed* the ghost. If you learn culture, it is because you were built to learn culture.

But in fact, we *don't* run into unmanageable meta-trouble in trying to come up with specific patches for specific known evolved biases. And evolution is stupid, so even though it has set up self-deceptive circuits in us, these circuits are not infinitely difficult to comprehend and outwit.

*Or so it seems!* But it really *does* seem that way, on reflection.

There is no button you can press to rewind past your noisy brain, and become a perfectly reliable ghost of perfect emptiness. That's not just because your brain *is you*. It's also because you can't unassume things like modus ponens<sup>15</sup> or belief updating<sup>16</sup>. You can unassume them as explicit premises for deliberate reasoning—a hunter-gatherer has no *explicit* concept of modus ponens—but you can't delete the actual dynamics (and all their products!).

So, in the end, I think we must allow the use of brains to think about thinking; and the use of evolved brains to think about evolution; and the use of inductive brains to think about induction; and the use of brains with an Occam prior to think about whether the universe appears to be simple; for these things we really *cannot* unwind entirely, even when we have reason to distrust them. Strange loops through the meta level, I think, are not the same as circular logic<sup>17</sup>.

---

13. [http://lesswrong.com/lw/kr/an\\_alien\\_god/](http://lesswrong.com/lw/kr/an_alien_god/)

14. [http://lesswrong.com/lw/kt/evolutions\\_are\\_stupid\\_but\\_work\\_anyway/](http://lesswrong.com/lw/kt/evolutions_are_stupid_but_work_anyway/)

15. Page 861, 'Created Already In Motion'.

16. Page 28, 'An Intuitive Explanation of Bayes' Theorem'.

17. Page 896, 'My Kind of Reflection'.



## 15. Rebelling Within Nature<sup>1</sup>

**Followup to:** Fundamental Doubts<sup>2</sup>, Where Recursive Justification Hits Bottom<sup>3</sup>, No Universally Compelling Arguments<sup>4</sup>, Joy in the Merely Real<sup>5</sup>, Evolutionary Psychology<sup>6</sup>

“Let us understand, once and for all, that the ethical progress of society depends, not on imitating the cosmic process, still less in running away from it, but in combating it.”

—T. H. Huxley (“Darwin’s bulldog”, early advocate of evolutionary theory)

There is a quote from some Zen Master<sup>7</sup> or other, who said something along the lines of:

“Western man believes that he is rebelling against nature, but he does not realize that, in doing so, he is acting according to nature.”

The Reductionist Masters of the West, strong in their own Art, are not so foolish; they *do* realize that they always act within Nature.

You can narrow your focus and rebel against a *facet* of existing Nature—polio, say—but in so doing, you act within the *whole* of Nature. The syringe that carries the polio vaccine is forged of atoms; our minds, that understood the method, embodied in neurons. If Jonas Salk had to fight laziness, he fought something that evolution instilled in him—a reluctance to work that conserves energy. And he fought it *with* other emotions that natural selection also inscribed in him: feelings of friendship that he extended to humanity, heroism to protect his tribe, maybe an explicit desire

---

1. [http://lesswrong.com/lw/sj/rebelling\\_within\\_nature/](http://lesswrong.com/lw/sj/rebelling_within_nature/)

2. Page 902, ‘Fundamental Doubts’.

3. Page 887, ‘Where Recursive Justification Hits Bottom’.

4. Page 844, ‘No Universally Compelling Arguments’.

5. Page 715, ‘Joy in the Merely Real’.

6. [http://lesswrong.com/lw/lx/evolutionary\\_psychology/](http://lesswrong.com/lw/lx/evolutionary_psychology/)

7. [http://lesswrong.com/lw/m7/zen\\_and\\_the\\_art\\_of\\_rationality/](http://lesswrong.com/lw/m7/zen_and_the_art_of_rationality/)

for fame that he never acknowledged to himself—who knows? (I haven't actually read a biography of Salk.)

The point is, you can't fight Nature from beyond Nature, only from within it. There is no acausal<sup>8</sup> fulcrum on which to stand outside reality and move it. There is no ghost of perfect emptiness<sup>9</sup> by which you can judge your brain from outside your brain. You can fight the cosmic process, but only by recruiting other abilities that evolution originally gave to you.

And if you fight one emotion within yourself—looking upon your own nature, and judging yourself less than you think should be—saying perhaps, “I should not *want* to kill my enemies”—then you make *that* judgment, by...

How exactly *does* one go about rebelling against one's own goal system?

From within it, naturally.

This is perhaps *the* primary thing that I didn't quite understand as a teenager<sup>10</sup>.

At the age of fifteen (fourteen?), I picked up a copy of TIME magazine and read an article on evolutionary psychology<sup>11</sup>. It seemed like one of the most massively obvious-in-retrospect ideas I'd ever heard. I went on to read *The Moral Animal* by Robert Wright. And later *The Adapted Mind*—but from the perspective of personal epiphanies, *The Moral Animal* pretty much did the job.

I'm reasonably sure that if I had not known the basics of evolutionary psychology from my teenage years, I would not currently exist as the Eliezer Yudkowsky you know.

Indeed, let me drop back a bit further:

At the age of... I think it was nine... I discovered the truth about sex by looking it up in my parents' home copy of the Encyclopedia Britannica (stop that laughing). Shortly after, I learned a good deal more by discovering where my parents had hidden the secret 15th volume of my long-beloved Childcraft series. I'd been avidly reading the first 14 volumes—some of them, anyway—since the age of

8. [http://lesswrong.com/lw/r/the\\_ultimate\\_source/](http://lesswrong.com/lw/r/the_ultimate_source/)

9. Page 844, ‘No Universally Compelling Arguments’.

10. Page 167, ‘My Wild and Reckless Youth’.

11. [http://lesswrong.com/lw/lx/evolutionary\\_psychology/](http://lesswrong.com/lw/lx/evolutionary_psychology/)

five. But the 15th volume wasn't meant for me—it was the “Guide for Parents”.

The 15th volume of Childcraft described the life cycle of children. It described the horrible confusion of the teenage years—teenagers experimenting with alcohol, with drugs, with unsafe sex, with reckless driving, the hormones taking over their minds, the overwhelming importance of peer pressure, the tearful accusations of “You don’t love me!” and “I hate you!”

I took one look at that description, at the tender age of nine, and said to myself in quiet revulsion, *I’m not going to do that.*

And I didn’t.

My teenage years were not untroubled. But I didn’t do any of the things that the *Guide to Parents* warned me against. I didn’t drink, drive, drug, lose control to hormones, pay any attention to peer pressure, or ever once think that my parents didn’t love me.

In a safer world, I would have wished for my parents to have hidden that book better.

But in this world, which needs me as I am, I don’t regret finding it.

I still rebelled, of course. I rebelled against the rebellious nature the *Guide to Parents* described to me. That was part of how I defined my identity in my teenage years—“I’m not doing the standard stupid stuff.” Some of the time, this just meant that I invented amazing new stupidity, but in fact that *was* a major improvement.

Years later, *The Moral Animal* made suddenly obvious the *why* of all that disastrous behavior I’d been warned against. Not that Robert Wright pointed any of this out explicitly, but it was obvious given the elementary concept of evolutionary psychology:

Physiologically adult humans are not meant to spend an additional 10 years in a school system; their brains map that onto “I have been assigned low tribal status”. And so, of course, they plot rebellion—accuse the existing tribal overlords of corruption—plot perhaps to split off their own little tribe in the savanna, not realizing that this is impossible in the Modern World. The teenage males map their own fathers onto the role of “tribal chief”...

Echoes in time, thousands of repeated generations in the savanna carving the pattern, ancient repetitions of form, reproduced in

the present in strange twisted mappings, across genes that didn't know anything had changed...

The world grew older, of a sudden.

And I'm not going to go into the evolutionary psychology of "teenagers" in detail, not now, because that would deserve its own post.

But when I read *The Moral Animal*, the world suddenly acquired *causal depth*. Human emotions existed for *reasons*, they weren't just unexamined givens. I might previously have questioned whether an emotion was appropriate to its circumstance—whether it made sense to hate your parents, if they did really love you—but I wouldn't have thought, before then, to judge *the existence of hatred as an evolved emotion*.

And then, having come so far, and having avoided with instinctive ease all the classic errors<sup>12</sup> that evolutionary psychologists are traditionally warned against—I was never once tempted to confuse evolutionary causation with psychological causation—I went wrong at the last turn.

The echo in time that was teenage psychology was obviously wrong and stupid—a *distortion* in the way things should be—so clearly you were supposed to unwind past it, compensate in the opposite direction or disable the feeling, to arrive at the correct answer.

It's hard for me to remember exactly what I was thinking in this era, but I think I tended to focus on one facet of human psychology at any given moment, trying to unwind myself a piece at a time. IIRC I did think, in full generality, "Evolution is bad; the effect it has on psychology is bad." (Like it had some kind of "effect" that could be isolated!) But somehow, I managed not to get to "Evolutionary psychology is the cause of altruism; altruism is bad."

It was easy for me to see all sorts of *warped* altruism as having been *warped by evolution*.

People who wanted to trust themselves with power, for the good of their tribe—that had an obvious evolutionary explanation; it was, therefore, a distortion to be corrected.

---

12. [http://lesswrong.com/lw/lx/evolutionary\\_psychology/](http://lesswrong.com/lw/lx/evolutionary_psychology/)

People who wanted to be altruistic in ways their friends would approve of—obvious evolutionary explanation; therefore a distortion to be corrected.

People who wanted to be altruistic in a way that would optimize their fame and repute—obvious evolutionary distortion to be corrected.

People who wanted to help only their family, or only their nation—acting out ancient selection pressures on the savanna; move past it.

But the fundamental will to help people?

Well, the notion of *that* being merely<sup>13</sup> evolved, was something that, somehow, I managed to *never quite accept*. Even though, in retrospect, the causality is just as obvious as teen revolutionism.

IIRC, I did think something along the lines of: “Once you unwind past evolution, then the true morality isn’t likely to contain a clause saying, ‘This person matters but this person doesn’t’, so everyone should matter equally, so you should be as eager to help others as help yourself.” And so I thought that even if the emotion of altruism had merely evolved, it was a right emotion, and I should keep it.

But why think that people mattered at all, if you were trying to unwind past all evolutionary psychology? Why think that it was better for people to be happy than sad, rather than the converse?

If I recall correctly, I *did* ask myself that, and sort of waved my hands mentally and said, “It just seems like one of the best guesses—I mean, I don’t know that people are valuable, but I can’t think of what else could be.”

This is the Avoiding Your Belief’s Real Weak Points<sup>14</sup> / Not Spontaneously Thinking About Your Belief’s Most Painful Weaknesses<sup>15</sup> antipattern in full glory: Get just far enough to place yourself on the first fringes of real distress, and then stop thinking.

---

13. Page 715, ‘Joy in the Merely Real’.

14. Page 525, ‘Avoiding Your Belief’s Real Weak Points’.

15. Page 525, ‘Avoiding Your Belief’s Real Weak Points’.

And also the antipattern of trying to unwind past everything<sup>16</sup> that is causally responsible for your existence as a mind<sup>17</sup>, to arrive at a perfectly reliable ghost of perfect emptiness<sup>18</sup>.

Later, having also seen others making similar mistakes, it seems to me that the general problem is an illusion of mind-independence that comes from picking something that appeals to you, while still seeming philosophically simple.

As if the appeal to you, of the moral argument, weren't still a feature of your particular point in mind design space<sup>19</sup>.

As if there weren't still an ordinary and explicable causal history<sup>20</sup> behind the appeal, and your selection of that particular principle.

As if, by making things philosophically simpler-seeming, you could enhance their appeal to a ghost-in-the-machine<sup>21</sup> who would hear your justifications starting from scratch, as fairness demands<sup>22</sup>.

As if your very sense of simplicity were not an aesthetic sense inscribed in you by evolution.

As if your very intuitions of “moral argument” and “justification”, were not an architecture-of-reasoning inscribed in you by natural selection, and just as causally explicable as any other feature of human psychology...

You can't throw away evolution, and end up with a perfectly moral creature that humans would have been, if only we had never evolved; that's really not how it works.

Why accept intuitively appealing arguments about the nature of morality, rather than intuitively unappealing ones, if you're going to distrust everything in you that ever evolved?

Then what *is* right? What *should* we do, having been inscribed by a blind mad idiot god<sup>23</sup> whose incarnation-into-reality takes the form of millions of years of ancestral murder and war?

But even this question—every fragment of it—the notion that a blind mad idiocy is an ugly property for a god to have, or that

16. Page 887, ‘Where Recursive Justification Hits Bottom’.

17. Page 902, ‘Fundamental Doubts’.

18. Page 844, ‘No Universally Compelling Arguments’.

19. [http://lesswrong.com/lw/rm/the\\_design\\_space\\_of\\_mindsgeneral/](http://lesswrong.com/lw/rm/the_design_space_of_mindsgeneral/)

20. [http://lesswrong.com/lw/rz/the\\_ultimate\\_source/](http://lesswrong.com/lw/rz/the_ultimate_source/)

21. [http://lesswrong.com/lw/rf/ghosts\\_in\\_the\\_machine/](http://lesswrong.com/lw/rf/ghosts_in_the_machine/)

22. Page 864, ‘The Bedrock of Fairness’.

23. [http://lesswrong.com/lw/kr/an\\_alien\\_god/](http://lesswrong.com/lw/kr/an_alien_god/)

murder is a poisoned well of order, even the words “right” and “should”—all a phenomenon within nature. All traceable back to debates built around arguments appealing to intuitions that evolved in me.

*You can't jump out of the system.* You really can't. Even *wanting* to jump out of the system—the sense that something isn't justified “just because it evolved”—is something that you feel from *within* the system. Anything you might try to use to jump—any sense of what morality *should* be like, if you could unwind past evolution—is *also* there as a causal result of evolution.

Not everything we think about morality is *directly* inscribed by evolution, of course. We have values that we got from our parents teaching them to us as we grew up; after it won out in a civilization-al debate conducted with reference to other moral principles; that were themselves argued into existence by appealing to built-in emotions; using an architecture-of-interpersonal-moral-argument that evolution burped into existence.

It all goes back to evolution. This doesn't just include things like instinctive concepts of fairness, or empathy, it includes the whole notion of arguing morals as if they were propositional beliefs. Evolution created within you that *frame of reference* within which you can *formulate the concept* of moral questioning. Including questioning evolution's fitness to create our moral frame of reference. If you *really* try to unwind outside the system, you'll unwind your unwinders.

That's what I didn't quite get, those years ago.

I do plan to dissolve the cognitive confusion that makes words like “right” and “should” seem difficult to grasp<sup>24</sup>. I've been working up to that for a while now.

But I'm not there yet, and so, for now, I'm going to jump ahead and peek at an answer I'll only later be able to justify as moral philosophy:

Embrace reflection<sup>25</sup>. You can't unwind to emptiness, but you can bootstrap from a starting point.

Go on morally questioning the existence (and not just appropriateness) of emotions. But don't treat the mere fact of their *having*

---

24. Page 879, ‘Is Morality Given?’.

25. Page 896, ‘My Kind of Reflection’.

*evolved* as a reason to reject them. Yes, I know that “X evolved” doesn’t seem like a good justification for having an emotion; but don’t let that be a reason to reject X, any more than it’s a reason to accept it. Hence the post on the Genetic Fallacy<sup>26</sup>: causation is conceptually distinct from justification. If you try to apply the Genetic Accusation to automatically convict and expel your *genes*, you’re going to run into foundational trouble<sup>27</sup>—so don’t!

Just ask if the emotion is justified—don’t treat its evolutionary cause as proof of mere distortion. Use your current mind<sup>28</sup> to examine the emotion’s pluses and minuses, without being ashamed; *use your full strength of morality*.

Judge emotions *as emotions*, not as evolutionary relics. When you say, “motherly love outcompeted its alternative alleles because it protected children that could carry the allele for motherly love”, this is only a *cause*, not a sum of all moral arguments. The evolutionary psychology may grant you helpful insight into the pattern and process of motherly love, but it neither justifies the emotion as natural, nor convicts it as coming from an unworthy source. You don’t make the Genetic Accusation either way. You just, y’know, think about motherly love, and ask yourself if it seems like a good thing or not; considering its effects, not its source.

You tot up the balance of moral justifications, using your current mind—with worrying about the fact that the entire debate takes place within an evolved framework.

That’s the moral normality<sup>29</sup> to which my yet-to-be-revealed moral philosophy will add up.

And if, in the meanwhile, it seems to you like I’ve just proved that there is no morality... well, I haven’t proved any such thing. But, meanwhile, just ask yourself if you might want to help people even if there were no morality<sup>30</sup>. If you find that the answer is yes, then you will later discover that you discovered morality.

---

26. Page 485, ‘The Genetic Fallacy’.

27. Page 902, ‘Fundamental Doubts’.

28. Page 887, ‘Where Recursive Justification Hits Bottom’.

29. Page 879, ‘Is Morality Given?’.

30. Page 856, ‘The Moral Void’.

## 16. Probability is Subjectively Objective<sup>1</sup>

**Followup to:** Probability is in the Mind<sup>2</sup>

“Reality is that which, when you stop believing in it,  
doesn’t go away.”

—Philip K. Dick

There are two kinds of Bayesians, allegedly. Subjective Bayesians believe that “probabilities” are degrees of uncertainty existing in our minds<sup>3</sup>; if you are uncertain about a phenomenon, that is a fact about your state of mind, not a property of the phenomenon itself; probability theory constrains the logical coherence of uncertain beliefs. Then there are objective Bayesians, who... I’m not quite sure what it means to be an “objective Bayesian”; there are multiple definitions out there. As best I can tell, an “objective Bayesian” is anyone who uses Bayesian methods and isn’t a subjective Bayesian.

If I recall correctly, E. T. Jaynes, master of the art, once described himself as a subjective-objective Bayesian. Jaynes certainly believed very firmly that probability was in the mind; Jaynes was the one who coined the term Mind Projection Fallacy<sup>4</sup>. But Jaynes also didn’t think that this implied a license to make up whatever priors you liked. There was only one *correct* prior distribution to use, given your state of partial information at the start of the problem.

How can something be in the mind, yet still be objective?

It appears to me that a good deal of philosophical maturity consists in being able to keep separate track of nearby concepts, without mixing them up<sup>5</sup>.

For example, to understand evolutionary psychology<sup>6</sup>, you have to keep separate track of the psychological purpose of an act, and the evolutionary pseudo-purposes of the adaptations that execute as the psychology; this is a common failure of newcomers to evo-

---

1. [http://lesswrong.com/lw/s6/probability\\_is\\_subjectively\\_objective/](http://lesswrong.com/lw/s6/probability_is_subjectively_objective/)

2. Page 687, ‘Probability is in the Mind’.

3. Page 687, ‘Probability is in the Mind’.

4. Page 687, ‘Probability is in the Mind’.

5. Page 253, ‘Fallacies of Compression’.

6. [http://lesswrong.com/lw/lx/evolutionary\\_psychology/](http://lesswrong.com/lw/lx/evolutionary_psychology/)

lutionary psychology, who read, misunderstand, and thereafter say, “You think you love your children, but you’re just trying to maximize your fitness!”

What is it, exactly, that the terms “subjective” and “objective”, mean? Let’s say that I hand you a sock. Is it a subjective or an objective sock? You believe that  $2 + 3 = 5$ . Is *your belief* subjective or objective? What about two plus three *actually* equaling five—is that subjective or objective? What about a specific act of adding two apples and three apples and getting five apples?

I don’t intend to confuse you in shrouds of words; but I do mean to point out that, while you may feel that you know very well what is “subjective” or “objective”, you might find that you have a bit of trouble saying out loud what those words mean.

Suppose there’s a calculator that computes “ $2 + 3 = 5$ ”. We punch in “2”, then “+”, then “3”, and lo and behold, we see “5” flash on the screen. We accept this as *evidence* that  $2 + 3 = 5$ , but we wouldn’t say that the calculator’s physical output *defines* the answer to the question  $2 + 3 = ?$ . A cosmic ray could strike a transistor, which might give us misleading evidence and cause us to believe that  $2 + 3 = 6$ , but it wouldn’t affect the *actual* sum of  $2 + 3$ .

Which proposition is common-sensically true, but philosophically interesting: while we can easily point to the physical location of a symbol on a calculator screen, or observe the result of putting two apples on a table followed by another three apples, it is rather harder to track down the whereabouts of  $2 + 3 = 5$ . (Did you look in the garage?)

But let us leave aside the question of *where* the fact  $2 + 3 = 5$  is located—in the universe, or somewhere else—and consider the assertion that the proposition is “objective”. If a cosmic ray strikes a calculator and makes it output “6” in response to the query “ $2 + 3 = ?$ ”, and you add two apples to a table followed by three apples, then you’ll still see five apples on the table. If you do the calculation in your own head, expending the necessary computing power—we assume that  $2 + 3$  is a very difficult sum to compute, so that the answer is not immediately obvious to you—then you’ll get the answer “5”. So the cosmic ray strike didn’t change anything.

---

<sup>7</sup>. Page 234, ‘Feel the Meaning’.

And similarly—exactly similarly<sup>8</sup>—what if a cosmic ray strikes a neuron inside your brain, causing you to compute “ $2 + 3 = 7$ ”? Then, adding two apples to three apples, you will expect to see seven apples, but instead you will be surprised to see five apples.

If instead we found that no one was ever mistaken about addition problems, and that, moreover, you could change the answer by an act of will, then we might be tempted to call addition “subjective” rather than “objective”. I am not saying that this is *everything* people mean by “subjective” and “objective”, just pointing to one aspect of the concept. One might summarize this aspect thus: “If you can change something by thinking differently, it’s subjective; if you can’t change it by anything you do strictly inside your head, it’s objective.”

Mind is not magic. Every act of reasoning that we human beings carry out, is *computed within* some particular human brain. But not every computation is *about* the state of a human brain. Not every thought that you think is *about* something that can be changed by thinking. Herein lies the opportunity for confusion-of-levels. The quotation is not the referent.<sup>9</sup> If you are going to consider thoughts as referential at all—if not, I’d like you to explain the mysterious correlation between my thought “ $2 + 3 = 5$ ” and the observed behavior of apples on tables—then, while the quoted thoughts will always change with thoughts, the referents *may or may not* be entities that change with changing human thoughts.

The calculator computes “What is  $2 + 3$ ”, not “What does this calculator compute as the result of  $2 + 3$ ?”. The answer to the former question is  $5$ , but if the calculator were to ask the latter question instead, the result could self-consistently be anything at all! If the calculator returned  $42$ , then indeed, “What does this calculator compute as the result of  $2 + 3$ ? ” would in fact be  $42$ .

So just because a computation takes place inside your brain, does not mean that the computation *explicitly mentions* your brain, that it has your brain as a *referent*, any more than the calculator mentions the calculator. The calculator does not attempt to contain a representation of itself, only of numbers.

---

8. [http://lesswrong.com/lw/k2/a\\_priori/](http://lesswrong.com/lw/k2/a_priori/)

9. Page 692, ‘The Quotation is not the Referent’.

Indeed, in the most straightforward implementation, the calculator that asks “What does this calculator compute as the answer to the query  $2 + 3 = ?$ ” will *never* return a result, just simulate itself simulating itself until it runs out of memory.

But if you punch the keys “2”, “+”, and “3”, and the calculator proceeds to compute “What do I output when someone punches ‘2 + 3?’”, the resulting computation does have one interesting characteristic: the *referent* of the computation is highly subjective, since it depends on the computation, and can be made to be anything just by changing the computation.

Is probability, then, subjective or objective?

Well, probability is computed within human brains or other calculators. A probability is a state of partial information that is possessed by you; if you flip a coin and press it to your arm, the coin is showing heads or tails, but you assign the probability  $1/2$  until you reveal it. A friend, who got a tiny but not fully informative peek, might assign a probability of 0.6.

So can you make the probability of winning the lottery be anything you like?

Forget about many-worlds for the moment—you should almost always be able to forget about many-worlds<sup>10</sup>—and pretend that you’re living in a single Small World where the lottery has only a single outcome. You will nonetheless have a need to call upon probability. Or if you prefer, we can discuss the ten trillionth decimal digit of pi, which I believe is not yet known. (If you are foolish enough to refuse to assign a probability distribution to this entity, you might pass up an excellent bet, like betting \$1 to win \$1000 that the digit is not 4.) Your uncertainty is a state of your mind, of partial information that you possess. Someone else might have different information, complete or partial. And the entity itself will only ever take on a single value.

So can you make the probability of winning the lottery, or the probability of the ten trillionth decimal digit of pi equaling 4, be anything you like?

You might be tempted to reply: “Well, since I *currently* think the probability of winning the lottery is one in a hundred million, then obviously, I will *currently* expect that assigning any other prob-

---

<sup>10</sup>. [http://lesswrong.com/lw/qz/living\\_in\\_many\\_worlds/](http://lesswrong.com/lw/qz/living_in_many_worlds/)

ability than this to the lottery, will decrease my expected log-score—or if you prefer a decision-theoretic formulation, I will expect this modification to myself to decrease expected utility. So, obviously, I will not choose to modify my probability distribution. It wouldn't be reflectively coherent."

So reflective coherency is the goal, is it? Too bad you weren't born with a prior that assigned probability 0.9 to winning the lottery! Then, by exactly the same line of argument, you wouldn't want to assign any probability except 0.9 to winning the lottery. And you would still be reflectively coherent. And you would have a 90% probability of winning millions of dollars! Hooray!

"No, then I would *think* I had a 90% probability of winning the lottery, but *actually*, the probability would only be one in a hundred million."

Well, of course *you* would be expected to say that. And if you'd been born with a prior that assigned 90% probability to your winning the lottery, you'd consider an alleged probability of  $10^{-8}$ , and say, "No, then I would *think* I had almost no probability of winning the lottery, but *actually*, the probability would be 0.9."

"Yeah? Then just modify your probability distribution, and buy a lottery ticket, and then wait and see what happens."

What happens? Either the ticket will win, or it won't. That's what will happen. We won't get to see that some particular probability was, in fact, the exactly right probability to assign.

"Perform the experiment a hundred times, and—"

Okay, let's talk about the ten trillionth digit of pi, then. Single-shot problem, no "long run" you can measure.

Probability is subjectively objective: Probability exists in your mind: if you're ignorant of a phenomenon, that's an attribute of you, not an attribute of the phenomenon. Yet it will seem to you that you can't change probabilities by wishing.

You could make yourself compute something *else*, perhaps, *rather than* probability. You could compute "What do I say is the probability?" (answer: anything you say) or "What do I wish were the probability?" (answer: whatever you wish) but these things are not the *probability*, which is subjectively objective.

The thing about subjectively objective quantities is that they *really do* seem objective to you. You don't look them over and say,

“Oh, well, of course I don’t want to modify my own probability estimate, because no one can just modify their probability estimate; but if I’d been born with a different prior I’d be saying something different, and I wouldn’t want to modify that either; and so none of us is superior to anyone else.” That’s the way a subjectively *subjective* quantity would seem.

No, it will seem to you that, if the lottery sells a hundred million tickets, and you don’t get a peek at the results, then the probability of a ticket winning, *is* one in a hundred million. And that you could be born with different priors but that wouldn’t give you any better odds. And if there’s someone next to you saying the same thing about *their* 90% probability estimate, you’ll just shrug and say, “Good luck with that.” You won’t expect them to *win*.

Probability is subjectively *really* objective, not just subjectively *sort of* objective.

Jaynes used to recommend that no one ever write out an unconditional probability: That you never, ever write simply  $P(A)$ , but always write  $P(A|I)$ , where  $I$  is your prior information. I’ll use  $Q$  instead of  $I$ , for ease of reading, but Jaynes used  $I$ . Similarly, one would not write  $P(A|B)$  for the posterior probability of  $A$  given that we learn  $B$ , but rather  $P(A|B,Q)$ , the probability of  $A$  given that we learn  $B$  and had background information  $Q$ .

This is good advice in a purely pragmatic sense, when you see how many false “paradoxes” are generated by accidentally using different prior information in different places.

But it also makes a deep philosophical point as well, which I never saw Jaynes spell out explicitly, but I think he would have approved: *there is no such thing as a probability that isn’t in any mind*. Any mind that takes in evidence and outputs probability estimates of the next event, remember, can be viewed as a prior<sup>ii</sup>—so there is no probability without priors/minds.

You can’t unwind the  $Q$ . You can’t ask “What is the *unconditional* probability of our background information being true,  $P(Q)$ ?” To make that estimate, you would still need *some* kind of prior. No way to unwind back to an ideal ghost of perfect emptiness...

You might argue that you and the lottery-ticket buyer do not really have a disagreement about *probability*. You say that the prob-

---

<sup>ii</sup>. [http://lesswrong.com/lw/hk/priors\\_as\\_mathematical\\_objects/](http://lesswrong.com/lw/hk/priors_as_mathematical_objects/)

ability of the ticket winning the lottery is one in a hundred million given your prior,  $P(W|Q_1) = 10^{-8}$ . The other fellow says the probability of the ticket winning given his prior is  $P(W|Q_2) = 0.9$ . Every time *you* say “The probability of X is Y”, you really mean, “ $P(X|Q_1) = Y$ ”. And when *he* says, “No, the probability of X is Z”, he really means, “ $P(X|Q_2) = Z$ ”.

Now you might, if you traced out his mathematical calculations, agree that, indeed, the conditional probability of the ticket winning, given his weird prior is 0.9. But you wouldn’t agree that “the probability of the ticket winning” is 0.9. Just as he wouldn’t agree that “the probability of the ticket winning” is  $10^{-8}$ .

Even if the two of you refer to different mathematical calculations when you say the word “probability”, *you* don’t think that puts you on equal ground, neither of you being better than the other. And neither does he, of course.

So you see that, subjectively, probability really *does* feel objective—even after you have subjectively taken all apparent subjectivity into account.

And this is not mistaken, because, by golly, the probability of winning the lottery really *is*  $10^{-8}$ , not 0.9. It’s not as if you’re doing your probability calculation *wrong*, after all. If you weren’t worried about being fair or about justifying yourself to philosophers, if you only wanted to get the correct answer<sup>12</sup>, your betting odds would be  $10^{-8}$ .

Somewhere out in mind design space<sup>13</sup>, there’s a mind with any possible prior; but that doesn’t mean that you’ll say, “All priors are created equal.”

When you judge those alternate minds, you’ll do so using your own mind—your own beliefs about the universe—your own posterior that came out of your own prior, your own posterior probability assignments  $P(X|A,B,C,\dots,Q_1)$ . But there’s nothing wrong with that<sup>14</sup>. It’s not like you could judge using something other than yourself. It’s not like you could have a probability assignment without any prior, a degree of uncertainty that isn’t in any mind.

---

12. Page 644, ‘Something to Protect’.

13. [http://lesswrong.com/lw/rm/the\\_design\\_space\\_of\\_minds\\_in\\_general/](http://lesswrong.com/lw/rm/the_design_space_of_minds_in_general/)

14. Page 896, ‘My Kind of Reflection’.

And so, when all that is said and done, it still seems like the probability of winning the lottery really *is*  $10^{-8}$ , not 0.9. No matter what other minds in design space say differently.

Which shouldn't be surprising. When you compute probabilities, you're thinking about lottery balls, not thinking about brains or mind designs or other people with different priors. Your probability computation makes no mention of that, any more than it explicitly represents itself. Your goal, after all, is to win, not to be fair. So of course probability will *seem* to be independent of what other minds might think of it.

Okay, but... you *still* can't win the lottery by assigning a higher probability to winning.

If you like, we could regard probability as an idealized computation, just like  $2 + 2 = 4$  seems to be independent of any particular error-prone calculator that computes it; and you could regard your mind as trying to approximate this ideal computation. In which case, it is good that your mind does not mention people's opinions, and only thinks of the lottery balls; the ideal computation makes no mention of people's opinions, and we are trying to reflect this ideal as accurately as possible...

But what you will calculate is the "ideal calculation" to plug into your betting odds, will depend on your prior, even though the calculation won't have an explicit dependency on "your prior". Someone who thought the universe was anti-Occamian, would advocate an anti-Occamian calculation, regardless of whether or not anyone thought the universe was anti-Occamian.

Your calculations get checked against reality, in a probabilistic way; you either win the lottery or not. But interpreting these results, is done with your prior; once again there is no probability that isn't in any mind.

I am not trying to argue that you can win the lottery by wishing, of course. Rather, I am trying to inculcate the ability to *distinguish between levels*.

When you think about the ontological nature of probability, and perform reductionism<sup>15</sup> on it—when you try to explain how "probability" fits into a universe in which states of mind do not exist *fundamentally*—then you find that probability is computed within a

---

<sup>15</sup>. Page 699, 'Reductionism'.

brain; and you find that other possible minds could perform mostly-analogous operations with different priors and arrive at different answers.

But, when you consider probability *as probability*, think about the *referent* instead of the thought process—which thinking you will do in your own thoughts, which are physical processes—then you will conclude that the vast majority of possible priors are *probably wrong*. (You will also be able to conceive of priors which are, in fact, better than yours, because they assign more probability to the actual outcome; you just won’t know in advance which alternative prior is the truly better one.)

If you again swap your goggles to think about how probability is implemented in the brain, the seeming objectivity of probability is the way the probability algorithm feels from inside<sup>16</sup>; so it’s no *mystery* that, considering probability as probability, you feel that it’s not subject to your whims. That’s just what the probability-computation would be expected to say, since the computation doesn’t represent any dependency on your whims.

But when you swap out those goggles and go back to thinking about probabilities, then, by golly, your algorithm seems to be *right* in computing that probability is not subject to your whims. You *can’t* win the lottery just by changing your beliefs about it. And if that is the way you would be expected to feel, then so what? The feeling has been explained, not explained away<sup>17</sup>; it is not a *mere* feeling. Just because a calculation is implemented in your brain, doesn’t mean it’s *wrong*, after all.

Your “probability that the ten trillionth decimal digit of pi is 4”, is an attribute of yourself, and exists in your mind; the real digit is either 4 or not. And if you could change your belief about the probability by editing your brain, you wouldn’t expect that to change the probability.

Therefore I say of probability that it is “subjectively objective”.

---

<sup>16</sup>. Page 223, ‘How An Algorithm Feels From Inside’.

<sup>17</sup>. Page 703, ‘Explaining vs. Explaining Away’.

## 17. Whither Moral Progress?<sup>1</sup>

### Followup to: Is Morality Preference?<sup>2</sup>

In the dialogue “Is Morality Preference?<sup>3</sup>”, Obert argues for the existence of moral progress by pointing to free speech, democracy, mass street protests against wars, the end of slavery... and we could also cite female suffrage, or the fact that burning a cat alive was once a popular entertainment... and many other things that our ancestors believed were right, but which we have come to see as wrong, or vice versa.

But Subhan points out that if your only measure of progress is to take a difference against your current state, then you can follow a random walk, and still see the appearance of inevitable progress.

One way of refuting the simplest version of this argument, would be to say that we don’t automatically think ourselves the very apex of possible morality; that we can imagine our descendants being more moral than us.

But can you *concretely* imagine a being morally wiser than yourself—one who knows that some particular thing is wrong, when you believe it to be right?

Certainly: I am not sure of the moral status of chimpanzees, and hence I find it easy to imagine that a future civilization will label them definitely people, and castigate us for failing to cryopreserve the chimpanzees who died in human custody.

Yet this still doesn’t prove the existence of moral progress. Maybe I am simply mistaken about the nature of changes in morality that have previously occurred—like looking at a time chart of “differences between past and present”, noting that the difference has been steadily decreasing, and saying, without being able to visualize it, “Extrapolating this chart into the future, we find that the future will be even less different from the present than the present.”

So let me throw the question open to my readers: Whither moral progress?

---

1. [http://lesswrong.com/lw/s9/whither\\_moral\\_progress/](http://lesswrong.com/lw/s9/whither_moral_progress/)

2. Page 871, ‘Is Morality Preference?’.

3. Page 871, ‘Is Morality Preference?’.

You might say, perhaps, “Over time, people have become more willing to help one another—that is the very substance and definition of moral progress.”

But as John McCarthy put it:

“If everyone were to live for others all the time, life would be like a procession of ants following each other around in a circle.”

Once you make “People helping each other more” the *definition* of moral progress, then people helping each other all the time, is *by definition* the *apex* of moral progress.

At the very least we have Moore’s Open Question: It is not clear that helping others all the time is *automatically* moral progress, whether or not you argue that it is; and so we apparently have some notion of what constitutes “moral progress” that goes beyond the direct identification with “helping others more often”.

Or if you identify moral progress with “Democracy!<sup>4</sup>”, then at some point there was a first democratic civilization—at some point, people went from having no notion of democracy as a good thing, to inventing the idea of democracy as a good thing. If increasing democracy is the very substance of moral progress, then how did this moral progress come about to exist in the world? How did people invent, without knowing it, this very substance of moral progress?

It’s easy to come up with *concrete* examples of moral progress. Just point to a moral disagreement between past and present civilizations; or point to a disagreement between yourself and present civilization, and claim that future civilizations might agree with you.

It’s harder to answer Subhan’s challenge—to show *directionality*, rather than a random walk, on the meta-level. And explain how this directionality is implemented, on the meta-level: how people go from not having a moral ideal, to having it.

(I have my own ideas about this, as some of you know. And I’ll thank you *not* to link to them in the comments, or quote them and attribute them to me, until at least 24 hours have passed from this post.)

---

4. Page 180, ‘Applause Lights’.



## 18. The Gift We Give To Tomorrow<sup>1</sup>

**Followup to:** Thou Art Godshatter<sup>2</sup>, Joy in the Merely Real<sup>3</sup>, Is Morality Given?<sup>4</sup>, Rebelling Within Nature<sup>5</sup> <sup>6</sup>

How, oh how, did an unloving and mindless universe, cough up minds who were capable of love?

“No mystery in that,” you say, “it’s just a matter of natural selection<sup>7</sup>. ”

But natural selection is cruel, bloody, and bloody stupid<sup>8</sup>. Even when, on the surface of things, biological organisms aren’t *directly* fighting each other—aren’t *directly* tearing at each other with claws—there’s still a deeper competition going on between the genes. Genetic information is created when genes increase their *relative* frequency in the next generation—what matters for “genetic fitness” is not how many children you have, but that you have *more* children than others. It is quite possible for a species to evolve to extinction<sup>9</sup>, if the winning genes are playing negative-sum games.

How, oh how, could such a process create beings capable of love?

“No mystery,” you say, “there is never any mystery-in-the-world; mystery is a property of questions, not answers<sup>10</sup>. A mother’s children share her genes, so the mother loves her children.”

But sometimes mothers adopt children, and still love them. And mothers love their children for themselves, not for their genes.

“No mystery,” you say, “Individual organisms are adaptation-executers, not fitness-maximizers<sup>11</sup>. Evolutionary psychology<sup>12</sup> is not about deliberately maximizing fitness—through most of human his-

---

1. [http://lesswrong.com/lw/sa/the\\_gift\\_we\\_give\\_to\\_tomorrow/](http://lesswrong.com/lw/sa/the_gift_we_give_to_tomorrow/)

2. [http://lesswrong.com/lw/l3/thou\\_art\\_godshatter/](http://lesswrong.com/lw/l3/thou_art_godshatter/)

3. Page 715, Joy in the Merely Real’.

4. Page 879, ‘Is Morality Given?’.

5. Page 907, ‘Rebelling Within Nature’.

6. [http://lesswrong.com/lw/l1/evolutionary\\_psychology/](http://lesswrong.com/lw/l1/evolutionary_psychology/)

7. [http://lesswrong.com/lw/kr/an\\_alien\\_god/](http://lesswrong.com/lw/kr/an_alien_god/)

8. [http://lesswrong.com/lw/kr/an\\_alien\\_god/](http://lesswrong.com/lw/kr/an_alien_god/)

9. [http://lesswrong.com/lw/l3/evolving\\_to\\_extinction/](http://lesswrong.com/lw/l3/evolving_to_extinction/)

10. Page 154, ‘Mysterious Answers to Mysterious Questions’.

11. [http://lesswrong.com/lw/lo/adaptationexecuters\\_not\\_fitnessmaximizers/](http://lesswrong.com/lw/lo/adaptationexecuters_not_fitnessmaximizers/)

12. [http://lesswrong.com/lw/lx/evolutionary\\_psychology/](http://lesswrong.com/lw/lx/evolutionary_psychology/)

tory, we didn't know genes existed. We don't calculate our acts' effect on genetic fitness consciously, or even subconsciously."

But human beings form friendships even with non-relatives: how, oh how, can it be?

"No mystery, for hunter-gatherers often play Iterated Prisoner's Dilemmas, the solution to which is reciprocal altruism. Sometimes the most dangerous human in the tribe is not the strongest, the prettiest, or even the smartest, but the one who has the most allies."

Yet not all friends are fair-weather friends; we have a concept of true friendship—and some people have sacrificed their life for their friends. Would not such a devotion tend to remove itself from the gene pool?

"You said it yourself: we have a concept of true friendship and fair-weather friendship. We can tell, or try to tell, the difference between someone who considers us a valuable ally, and someone executing the friendship adaptation. We wouldn't be true friends with someone who we didn't think was a true friend to us—and someone with many *true* friends is far more formidable than someone with many fair-weather allies."

And Mohandas Gandhi, who really did turn the other cheek? Those who try to serve all humanity, whether or not all humanity serves them in turn?

"That perhaps is a more complicated story. Human beings are not just social animals. We are political animals who argue linguistically about policy in adaptive tribal contexts. Sometimes the formidable human is not the strongest, but the one who can most skillfully argue that their preferred policies match the preferences of others."

Um... that doesn't explain Gandhi, or am I missing something?

"The point is that we have the ability to *argue* about 'What should be done?' as a *proposition*—we can make those arguments and respond to those arguments, without which politics could not take place."

Okay, but Gandhi?

"Believed certain complicated propositions about 'What should be done?' and did them."

That sounds like it could explain any possible<sup>13</sup> human behavior.

“If we traced back the chain of causality through all the arguments, it would involve: a moral architecture that had the ability to argue *general abstract* moral propositions like ‘What should be done to people?'; appeal to hardwired intuitions like fairness, a concept of duty, pain aversion + empathy; something like a preference for simple moral propositions, probably reused from our previous Occam prior; and the end result of all this, plus perhaps memetic selection effects, was ‘You should not hurt people' in full generality—”

And that gets you Gandhi.

“Unless you think it was magic, it has to fit into the lawful causal development of the universe somehow.”

Well... I certainly won't postulate magic, under any name<sup>14</sup>.

“Good.”

But come on... doesn't it seem a little... *amazing*... that hundreds of millions of years worth of evolution's death tournament could cough up mothers and fathers, sisters and brothers, husbands and wives, steadfast friends and honorable enemies, true altruists and guardians of causes, police officers and loyal defenders, even artists sacrificing themselves for their art, all practicing so many kinds of love? For so many things other than genes<sup>15</sup>? Doing their part to make their world less ugly, something besides a sea of blood and violence and mindless replication?

“Are you claiming to be surprised by this? If so, question your underlying model, for it has led you to be surprised by the true state of affairs<sup>16</sup>. Since the beginning, not one unusual thing has ever happened.”

But how is it *not* surprising?

“What are you suggesting, that some sort of shadowy figure stood behind the scenes and directed evolution?”

Hell no. But—

“Because if you *were* suggesting that, I would have to ask how that shadowy figure *originally* decided that love was a *desirable* outcome of evolution. I would have to ask where that figure got

13. Page 140, ‘Guessing the Teacher's Password’.

14. Page 158, ‘The Futility of Emergence’.

15. [http://lesswrong.com/lw/l3/thou\\_art\\_godsbatter/](http://lesswrong.com/lw/l3/thou_art_godsbatter/)

16. [http://lesswrong.com/lw/hs/think\\_like\\_reality/](http://lesswrong.com/lw/hs/think_like_reality/)

preferences that included things like love, friendship, loyalty, fairness, honor, romance, and so on. On evolutionary psychology, we can see how *that specific outcome* came about—how *those particular goals rather than others* were generated in the first place. You can call it ‘surprising’ all you like. But when you really do understand evolutionary psychology, you can see how parental love and romance and honor, and even true altruism and moral arguments, bear the specific design signature of natural selection in particular adaptive contexts of the hunter-gatherer savanna. So if there was a shadowy figure, it must itself have evolved—and that obviates the whole point of postulating it.”

I’m not postulating a shadowy figure! I’m just asking how human beings ended up so nice.

“Nice! Have you *looked* at this planet lately? We also bear all those other emotions that evolved, too—which would tell you very well that we evolved, should you begin to doubt it. Humans aren’t always nice.”

We’re one hell of a lot nicer than the process that produced us, which lets elephants starve to death when they run out of teeth, and doesn’t anesthetize a gazelle even as it lays dying and is of no further importance to evolution one way or the other. It doesn’t take much to be nicer than evolution. To have the *theoretical capacity* to make one single gesture of mercy, to feel a single twinge of empathy, is to be nicer than evolution. How did evolution, which is itself so uncaring, create minds on that qualitatively higher moral level than itself? How did evolution, which is so ugly, end up doing anything so beautiful?

“Beautiful, you say? Bach’s *Little Fugue in G Minor* may be beautiful, but the sound waves, as they travel through the air, are not stamped with tiny tags to specify their beauty. If you wish to find explicitly encoded a measure of the fugue’s beauty, you will have to look at a human brain—nowhere else in the universe will you find it. Not upon the seas or the mountains will you find such judgments written: they are not minds, they cannot think.”

Perhaps that is so, but still I ask: How did evolution end up doing anything so beautiful, as giving us the ability to admire the beauty of a flower?

“Can you not see the circularity in your question? If beauty were like some great light in the sky that shined from outside humans, then your question might make sense—though there would still be the question of how humans came to perceive that light. You evolved with a psychology unlike evolution: Evolution has nothing like the intelligence or the precision required to exactly quine its goal system. In coughing up the first true minds, evolution’s simple fitness criterion shattered into a thousand values<sup>17</sup>. You evolved with a psychology that attaches utility<sup>18</sup> to things which evolution does not care about, like human life and happiness. And then you look back and say, ‘How marvelous, that uncaring evolution produced minds that care about sentient life!’ So your great marvel and wonder, that seems like far too much coincidence, is really no coincidence at all.”

But then it is still amazing that this particular circular loop, happened to loop around such important things as beauty and altruism.

“I don’t think you’re following me here. To you, it seems natural to privilege the beauty and altruism as special, as preferred, because you value them highly; and you don’t see this as a unusual fact about yourself, because many of your friends do likewise. So you expect that a ghost of perfect emptiness<sup>19</sup> would also value life and happiness—and then, from this standpoint outside reality, a great coincidence would indeed have occurred.”

But you can make arguments for the importance of beauty and altruism from first principles—that our aesthetic senses lead us to create new complexity, instead of repeating the same things over and over; and that altruism is important because it takes us outside ourselves, gives our life a higher meaning than sheer brute selfishness.

“Oh, and *that* argument is going to move even a ghost of perfect emptiness<sup>20</sup>—now that you’ve appealed to slightly different values? Those aren’t first principles, they’re just *different* principles. Even if you’ve adopted a high-falutin’ philosophical tone, still there are no *universally* compelling arguments. All you’ve done is pass the recursive buck<sup>21</sup>.”

---

<sup>17</sup>. [http://lesswrong.com/lw/l3/thou\\_art\\_godshatter/](http://lesswrong.com/lw/l3/thou_art_godshatter/)

<sup>18</sup>. [http://lesswrong.com/lw/l4/terminal\\_values\\_and\\_instrumental\\_values/](http://lesswrong.com/lw/l4/terminal_values_and_instrumental_values/)

<sup>19</sup>. Page 844, ‘No Universally Compelling Arguments’.

<sup>20</sup>. Page 844, ‘No Universally Compelling Arguments’.

You don't think that, somehow, we evolved to *tap into* something beyond—

"What good does it do to suppose something beyond? Why should we pay more attention to that beyond thing, than we pay to our existence as humans? How does it alter your personal responsibility, to say that you were only following the orders of the beyond thing? And you would still have evolved to let the beyond thing, rather than something else, direct your actions. You are only passing the recursive buck<sup>22</sup>. Above all, it would be *too much coincidence*."

Too much coincidence?

"A flower is beautiful, you say. Do you think there is no story behind that beauty, or that science does not know the story? Flower pollen is transmitted by bees, so by sexual selection, flowers evolved to attract bees—by imitating certain mating signs of bees, as it happened; the flowers' patterns would look more intricate, if you could see in the ultraviolet. Now healthy flowers are a sign of fertile land, likely to bear fruits and other treasures, and probably prey animals as well; so is it any wonder that humans evolved to be attracted to flowers? But for there to be some great light written upon the very stars—those huge unsentient balls of burning hydrogen—which *also* said that flowers were beautiful, now *that* would be far too much coincidence."

So you explain away<sup>23</sup> the beauty of a flower?

"No, I explain it. Of course there's a story behind the beauty of flowers and the fact that we find them beautiful. Behind ordered events, one finds ordered stories; and what has no story is the product of random noise, which is hardly any better. If you cannot take joy in things that have stories behind them, your life will be empty indeed.<sup>24</sup> I don't think I take any less joy in a flower than you do; more so, perhaps, because I take joy in its story as well."

Perhaps as you say, there is no surprise from a causal viewpoint—no disruption of the physical order of the universe. But it still seems to me that, in this creation of humans by evolution, something happened that is precious and marvelous and wonderful. If we cannot call it a physical miracle, then call it a moral miracle.

---

21. [http://lesswrong.com/lw/rd/passing\\_the\\_recursive\\_buck/](http://lesswrong.com/lw/rd/passing_the_recursive_buck/)

22. [http://lesswrong.com/lw/rd/passing\\_the\\_recursive\\_buck/](http://lesswrong.com/lw/rd/passing_the_recursive_buck/)

23. Page 703, 'Explaining vs. Explaining Away'.

24. Page 715, 'Joy in the Merely Real'.

“Because it’s only a miracle from the perspective of the morality that was produced, thus explaining away all of the apparent coincidence from a merely causal and physical perspective?”

Well... I suppose you could interpret the term that way, yes. I just meant something that was immensely surprising and wonderful on a moral level, even if it is not surprising on a physical level.

“I think that’s what I said.”

But it still seems to me that you, from your own view, drain something of that wonder away.

“Then you have problems taking joy in the merely real<sup>25</sup>. Love has to begin *somewhere*, it has to enter the universe *somewhere*. It is like asking how life itself begins—and though you were born of your father and mother, and they arose from their living parents in turn, if you go far and far and far away back, you will finally come to a replicator that arose by pure accident—the border between life and unlife. So too with love.

“A complex pattern must be explained by a cause which is not already that complex pattern. Not just the event must be explained, but the very shape and form. For love to first enter Time, it must come of something that is not love; if this were not possible, then love could not be.

“Even as life itself required that first replicator to come about by accident, parentless but still caused: far, far back in the causal chain that led to you: 3.85 billion years ago, in some little tidal pool.

“Perhaps your children’s children will ask how it is that they are capable of love.

“And their parents will say: Because we, who also love, created you to love.

“And your children’s children will ask: But how is it that *you* love?

“And their parents will reply: Because our own parents, who also loved, created us to love in turn.

“Then your children’s children will ask: But where did it all begin? Where does the recursion end?

“And their parents will say: Once upon a time, long ago and far away, ever so long ago, there were intelligent beings who were not

---

<sup>25</sup>. Page 715, ‘Joy in the Merely Real’.

themselves intelligently designed. Once upon a time, there were lovers created by something that did not love.

“Once upon a time, when all of civilization was a single galaxy and a single star: and a single planet, a place called Earth.

“Long ago, and far away, ever so long ago.”

## 19. Could Anything Be Right?<sup>1</sup>

**Followup to:** Where Recursive Justification Hits Bottom<sup>2</sup>, Rebelling Within Nature<sup>3</sup>

Years ago<sup>4</sup>, Eliezer<sub>1999</sub> was convinced that he knew *nothing* about morality.

For all he knew, morality could require the extermination of the human species; and if so he saw no virtue in taking a stand against morality<sup>5</sup>, because he thought that, by definition<sup>6</sup>, if he postulated that moral fact, that meant human extinction was what “should” be done.

I thought I could *figure out* what was right, perhaps, given enough reasoning time and enough facts, but that I currently had no information about it. I could not trust evolution which had built me<sup>7</sup>. What foundation<sup>8</sup> did that leave on which to stand?

Well, indeed Eliezer<sub>1999</sub> was massively mistaken about the nature of morality, so far as his explicitly represented philosophy went.

But as Davidson once observed, 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. You must get at least some of your beliefs right, before the remaining ones can be wrong *about* anything.

My belief that I had *no* information *about* morality was not internally consistent.

Saying that I knew nothing felt virtuous, for I had once been taught that it was virtuous to confess my ignorance<sup>9</sup>. “The only thing I know is that I know nothing,” and all that. But in this case I would have been better off considering the admittedly exaggerated saying, “The greatest fool is the one who is not aware they are

---

1. <http://lesswrong.com/lw/sb/couldAnythingBeRight/>

2. Page 887, ‘Where Recursive Justification Hits Bottom’.

3. Page 907, ‘Rebelling Within Nature’.

4. Page 167, ‘My Wild and Reckless Youth’.

5. Page 856, ‘The Moral Void’.

6. Page 193, ‘The Parable of Hemlock’.

7. Page 907, ‘Rebelling Within Nature’.

8. Page 887, ‘Where Recursive Justification Hits Bottom’.

9. Page 585, ‘The Proper Use of Humility’.

wise.” (This is nowhere near the *greatest* kind of foolishness, but it is a kind of foolishness.)

Was it wrong to kill people? Well, I thought so, but I wasn’t sure; maybe it was right to kill people, though that seemed less likely.

What kind of *procedure* would answer whether it was right to kill people? I didn’t know that either, but I thought that if you built a generic superintelligence (what I would later label a “ghost of perfect emptiness<sup>10</sup>”) then it could, you know, reason about what was likely to be right and wrong; and since it was *superintelligent*, it was bound to come up with the right answer.

The problem that I somehow managed not to think too hard about<sup>11</sup>, was where the superintelligence would get the procedure that discovered the procedure that discovered the procedure that discovered morality—if I couldn’t write it into the start state that wrote the successor AI that wrote the successor AI.

As Marcello Herreshoff later put it, “We never bother running a computer program unless we don’t know the output and we know an important fact about the output.” If I knew nothing about morality, and did not even claim to know the nature of morality, then how could I construct any computer program whatsoever—even a “superintelligent” one or a “self-improving” one—and claim that it would output something called “morality”?

There are no-free-lunch theorems in computer science—in a maxentropy universe, no plan is better on average than any other. If you have no knowledge at all about “morality”, there’s also no computational procedure that will seem more likely than others to compute “morality”, and no meta-procedure that’s more likely than others to produce a procedure that computes “morality”.

I thought that surely even a ghost of perfect emptiness, finding that it knew nothing of morality, would see a moral imperative to *think about morality*.

But the difficulty lies in the word *think*. Thinking is not an activity that a ghost of perfect emptiness is automatically able to carry out. Thinking requires running some *specific* computation that is the thought. For a reflective AI to decide to think, requires that

---

<sup>10</sup>. Page 844, ‘No Universally Compelling Arguments’.

<sup>11</sup>. Page 525, ‘Avoiding Your Belief’s Real Weak Points’.

it know some computation which it believes is *more* likely to tell it what it wants to know, than consulting an Ouija board; the AI must also have a notion of how to interpret the output.

If one knows nothing about morality, what does the word “should” mean, at all? If you don’t know whether death is right or wrong—and don’t know how you can discover whether death is right or wrong—and don’t know whether any given procedure might *output* the procedure for saying whether death is right or wrong—then what do these words, “right” and “wrong”, even *mean*?

If the words “right” and “wrong” have *nothing* baked into them—no starting point—if *everything* about morality is up for grabs, not just the content but the structure and the starting point and the determination procedure—then what is their meaning? What distinguishes, “I don’t know what is right” from “I don’t know what is wakalixes”?

A scientist may say that everything is up for grabs in science, since any theory may be disproven; but then they have some idea of what would count as *evidence* that could disprove the theory. Could there be something that would change what a scientist regarded as evidence?

Well, yes, in fact; a scientist who read some Karl Popper and thought they knew what “evidence” meant, could be presented with the coherence and uniqueness proofs underlying Bayesian probability, and that might change their definition of evidence. They might not have had any *explicit notion*, in advance, that such a proof could exist. But they would have had an *implicit notion*. It would have been baked into their brains, if not explicitly represented therein, that such-and-such an argument would in fact persuade them that Bayesian probability gave a better definition of “evidence” than the one they had been using.

In the same way, you could say, “I don’t know what morality is, but I’ll know it when I see it,” and make sense.

But then you are not rebelling completely against your own evolved nature<sup>12</sup>. You are supposing that whatever has been baked into you to recognize “morality”, is, if not absolutely trustworthy, then at least your initial condition with which you start debating. Can you trust your moral intuitions to give you any information

---

<sup>12</sup>. Page 907, ‘Rebelling Within Nature’.

about morality *at all*, when they are the product of mere evolution<sup>13</sup>?

But if you discard every procedure that evolution gave you *and all its products*, then you discard your whole brain. You discard everything that could potentially recognize morality when it sees it. You discard everything that could potentially respond to moral arguments by updating your morality. You even unwind past the unwinder: you discard the intuitions underlying your conclusion that *you can't trust evolution* to be moral. It is your *existing* moral intuitions that tell you that evolution doesn't seem like a very *good* source of morality. What, then, will the words "right" and "should" and "better" even *mean*?

Humans do not perfectly recognize truth when they see it, and hunter-gatherers do not have an explicit concept of the Bayesian criterion of evidence. But all our science and all our probability theory was built on top of a chain of appeals to our instinctive notion of "truth". Had this core been flawed, there would have been nothing we could do *in principle* to arrive at the present notion of science; the notion of science would have just sounded completely unappealing and pointless.

One of the arguments that might have shaken my teenage self out of his mistake, if I could have gone back in time to argue with him, was the question:

Could there be some morality, some given rightness or wrongness, that human beings do not perceive, do not want to perceive, will not see any appealing moral argument for adopting, nor any moral argument for adopting a procedure that adopts it, etcetera? Could there be a morality, and ourselves *utterly* outside its frame of reference? But then what makes this thing *morality*—rather than a stone tablet somewhere with the words 'Thou shalt murder' written on them, with absolutely no *justification* offered?

So all this suggests that you should be willing to accept that you might know a *little* about morality. Nothing unquestionable, perhaps, but an initial state with which to start questioning yourself<sup>14</sup>. Baked into your brain but not explicitly known to you, perhaps; but still, that which your brain *would* recognize as *right* is what you are

---

<sup>13</sup>. Page 927, 'The Gift We Give To Tomorrow'.

<sup>14</sup>. Page 887, 'Where Recursive Justification Hits Bottom'.

talking *about*. You will accept at least enough of the way you *respond to moral arguments* as a *starting point*, to identify “morality” as something to think about.

But that’s a rather large step.

It implies accepting your own mind as identifying a moral frame of reference, rather than all morality being a great light shining from beyond (that in principle you might not be able to perceive at all). It implies accepting that even if there were a light and your brain decided to recognize it as “morality”, it would still be your own brain that recognized it, and you would not have evaded causal responsibility—or evaded moral responsibility either, on my view.

It implies dropping the notion that a ghost of perfect emptiness will necessarily agree with you, because the ghost might occupy a different moral frame of reference, respond to different arguments, be *asking a different question* when it computes what-to-do-next.

And if you’re willing to bake at least a few things into the very meaning of this topic of “morality”, this quality of *rightness* that you are talking about when you talk about “rightness”—if you’re willing to accept even that morality is what you argue about when you argue about “morality”—then why not accept other intuitions, other pieces of yourself, into the starting point as well?

Why not accept that, *ceteris paribus*, joy is preferable to sorrow?

You might later find some ground within yourself or built upon yourself with which to criticize this—but why not accept it for now? Not just as a personal preference, mind you; but as something baked into the *question* you ask when you ask “What is truly right”?

But then you might find that you know rather a lot about morality! Nothing certain—nothing unquestionable—nothing unarguable—but still, quite a bit of information. Are you willing to relinquish your Socratean ignorance?

I don’t argue by definitions<sup>15</sup>, of course. But if you claim to know nothing at all about morality, then you will have problems with the *meaning* of your words, not just their plausibility<sup>16</sup>.

---

<sup>15</sup>. Page 193, ‘The Parable of Hemlock’.

<sup>16</sup>. Page 799, ‘The Generalized Anti-Zombie Principle’.

## 20. Existential Angst Factory<sup>1</sup>

### Followup to: The Moral Void<sup>2</sup>

A widespread excuse for avoiding rationality is the widespread belief that it is “rational” to believe life is meaningless, and thus suffer existential angst. This is one of the secondary reasons why it is worth discussing the nature of morality. But it’s also worth attacking existential angst directly.

I suspect that most existential angst is not really existential. I think that most of what is labeled “existential angst” comes from trying to solve the wrong problem<sup>3</sup>.

Let’s say you’re trapped in an unsatisfying relationship, so you’re unhappy. You consider going on a skiing trip, or you actually go on a skiing trip, and you’re still unhappy. You eat some chocolate, but you’re still unhappy. You do some volunteer work at a charity (or better yet, work the same hours professionally and donate the money<sup>4</sup>, thus applying the Law of Comparative Advantage) and you’re still unhappy because you’re in an unsatisfying relationship.

So you say something like: “Skiing is meaningless, chocolate is meaningless, charity is meaningless, life is doomed to be an endless stream of woe.” And you blame this on the universe being a mere dance of atoms, empty of meaning. Not necessarily because of some kind of subconsciously deliberate Freudian substitution to avoid acknowledging your real problem, but because you’ve stopped hoping that your real problem is solvable. And so, as a sheer unexplained background fact, you observe that you’re always unhappy.

Maybe you’re poor, and so always unhappy. Nothing you do solves your poverty, so it starts to seem like a universal background fact, along with your unhappiness. So when you *observe* that you’re always unhappy, you blame this on the universe being a mere dance of atoms. Not as some kind of Freudian substitution, but because it has *ceased to occur to you* that there *does* exist some possible state of affairs in which life is not painful.

---

1. [http://lesswrong.com/lw/sc/existential\\_angst\\_factory/](http://lesswrong.com/lw/sc/existential_angst_factory/)

2. Page 856, ‘The Moral Void’.

3. <http://cestelena.wordpress.com/2008/03/04/youre-solving-the-wrong-problem/>

4. [http://lesswrong.com/lw/65/money\\_the\\_unit\\_of\\_caring/](http://lesswrong.com/lw/65/money_the_unit_of_caring/)

What about rich heiresses with everything in the world available to buy, who still feel unhappy? Perhaps they can't get themselves into satisfying romantic relationships. One way or another, they don't *know* how to use their money to create happiness—they lack the expertise in hedonic psychology and/or self-awareness and/or simple competence.

So they're constantly unhappy—and they blame it on existential angst, because they've already solved the only problem they know how to solve. They already have enough money and they've already bought all the toys. Clearly, if there's still a problem, it's because life is meaningless.

If someone who weighs 560 pounds suffers from “existential angst”, *allegedly* because the universe is a mere dance of particles<sup>5</sup>, then stomach reduction surgery might drastically change their views of the metaphysics of morality.

I'm not a fan of Timothy Ferris, but *The Four-Hour Workweek* does make an interesting fun-theoretic<sup>6</sup> observation:

Let's assume we have 10 goals and we achieve them—what is the desired outcome that makes all the effort worthwhile? The most common response is what I also would have suggested five years ago: happiness. I no longer believe this is a good answer. Happiness can be bought with a bottle of wine and has become ambiguous through overuse. There is a more precise alternative that reflects what I believe the actual objective is.

Bear with me. What is the opposite of happiness? Sadness? No. Just as love and hate are two sides of the same coin, so are happiness and sadness. Crying out of happiness is a perfect illustration of this. The opposite of love is indifference, and the opposite of happiness is—here's the clincher—boredom.

*Excitement is the more practical synonym for happiness, and it is precisely what you should strive to chase. It is the cure-all.*

When people suggest you follow your “passion” or your

---

5. Page 715, ‘Joy in the Merely Real’.

6. [http://lesswrong.com/lw/xy/the\\_fun\\_theory\\_sequence/](http://lesswrong.com/lw/xy/the_fun_theory_sequence/)

“bliss,” I propose that they are, in fact, referring to the same singular concept: excitement.

This brings us full circle. The question you should be asking isn’t “What do I want?” or “What are my goals?” but “What would excite me?”

Remember—boredom is the enemy, not some abstract “failure.”

*Living like a millionaire requires doing interesting things and not just owning enviable things.*

I don’t endorse all of the above, of course. But note the SolvingTheWrongProblem<sup>7</sup> anti-pattern Ferris describes. It was on reading the above that I first generalized ExistentialAngstFactory.

Now, *if* someone is in a unproblematic, loving relationship; and they have enough money; and no major health problems; and they’re signed up for cryonics so death is not approaching inexorably; and they’re doing exciting work that they enjoy; and they believe they’re having a positive effect on the world...

...and they’re *still* unhappy because it seems to them that the universe is a mere dance of atoms empty of meaning, *then* we may have a legitimate problem here. One that, perhaps, can *only* be resolved by a very long discussion of the nature of morality and how it fits into a reductionist universe<sup>8</sup>.

But, mostly, I suspect that when people complain about the empty meaningless void, it is because they have at least one problem that they aren’t thinking about solving—perhaps because they never identified it. Being able to identify your own problems is a feat of rationality that schools don’t explicitly train you to perform. And they haven’t even been told that an un-focused-on problem might be the source of their “existential angst”—they’ve just been told to blame it on existential angst.

That’s the other reason it might be helpful to understand the nature of morality—even if it just adds up to moral normality—because it tells you that if you’re constantly unhappy, it’s *not* because the universe is empty of meaning.

---

7. <http://cestelena.wordpress.com/2008/03/04/youre-solving-the-wrong-problem/>

8. [http://wiki.lesswrong.com/wiki/Metaethics\\_sequence](http://wiki.lesswrong.com/wiki/Metaethics_sequence)

Or maybe believing the universe is a “mere dance of particles” is one more factor contributing to human unhappiness; in which case, again, people can benefit from eliminating that factor.

If it seems to you like nothing you do makes you happy, and you can’t even imagine what would make you happy, it’s not because the universe is made of particle fields. It’s because you’re *still* solving the wrong problem. Keep searching, until you find the visualizable state of affairs in which the existential angst seems like it should go away—that might (or might not) tell you the real problem; but at least, don’t blame it on reductionism.

**Added:** Several commenters pointed out that random acts of brain chemistry may also be responsible for depression, even if your life is otherwise fine. As far as I know, this is true. But, once again, it won’t help to mistake that random act of brain chemistry as being *about* existential issues; that might prevent you from trying pharmaceutical interventions.

## 21. Can Counterfactuals Be True?<sup>1</sup>

**Followup to:** Probability is Subjectively Objective<sup>2</sup>

The classic explanation of counterfactuals begins with this distinction:

1. If Lee Harvey Oswald didn't shoot John F. Kennedy, then someone else did.
2. If Lee Harvey Oswald hadn't shot John F. Kennedy, someone else would have.

In ordinary usage we would agree with the first statement, but not the second (I hope).

If, somehow, we learn the definite fact that Oswald did not shoot Kennedy, then someone else must have done so, since Kennedy was in fact shot.

But if we went back in time and removed Oswald, while leaving everything else the same, then—unless you believe there was a conspiracy<sup>3</sup>—there's no particular reason to believe Kennedy would be shot:

We start by imagining the same historical situation that existed in 1963—by a further act of imagination, we remove Oswald from our vision—we run forward the laws that we think govern the world—visualize Kennedy parading through in his limousine—and find that, in our imagination, no one shoots Kennedy.

It's an interesting question whether counterfactuals can be *true* or *false*. We never get to experience them directly.

If we disagree on what *would have* happened if Oswald hadn't been there, what experiment<sup>4</sup> could we perform to find out which of us is right?

And if the counterfactual is something unphysical—like, “If gravity had stopped working three days ago, the Sun would have exploded”—then there aren't even any alternate histories<sup>5</sup> out there to provide a truth-value.

---

1. [http://lesswrong.com/lw/sh/can\\_counterfactuals\\_be\\_true/](http://lesswrong.com/lw/sh/can_counterfactuals_be_true/)

2. Page 915, ‘Probability is Subjectively Objective’.

3. <http://www.theonion.com/content/node/42834>

4. Page 105, ‘Making Beliefs Pay Rent (in Anticipated Experiences)’.

5. [http://lesswrong.com/lw/r8/and\\_the\\_winner\\_is\\_manyworlds/](http://lesswrong.com/lw/r8/and_the_winner_is_manyworlds/)

It's not as simple as saying that if the bucket contains three pebbles, and the pasture contains three sheep, the bucket is true<sup>6</sup>.

Since the counterfactual event *only* exists in your imagination, how can it be true or false?

So... is it just as fair to say that “If Oswald hadn’t shot Kennedy, the Sun would have exploded”?

After all, the event only exists in our imaginations—surely that means it’s subjective, so we can say anything we like<sup>7</sup>?

But so long as we have a lawful specification of how counterfactuals are constructed—a lawful computational procedure—then the counterfactual result of removing Oswald, depends entirely on the empirical state of the world.

If there was no conspiracy, then any reasonable computational procedure that simulates removing Oswald’s bullet from the course of history, ought to return an answer of Kennedy not getting shot.

“Reasonable!” you say. “Ought!” you say.

But that’s not the point; the point is that if you *do* pick some fixed computational procedure, whether it is reasonable or not, then either it *will* say that Kennedy gets shot, or not, and what it says will depend on the empirical state of the world. So that, if you tell me, “I believe that *this-and-such* counterfactual construal, run over Oswald’s removal, preserves Kennedy’s life”, then I can deduce that you don’t believe in the conspiracy.

Indeed, so long as we take this computational procedure as fixed, then the actual state of the world (which either does include a conspiracy, or does not) presents a ready truth-value for the output of the counterfactual.

In general, if you give me a fixed computational procedure, like “multiply by 7 and add 5”, and then you point to a 6-sided die underneath a cup, and say, “The result-of-procedure is 26!” then it’s not hard at all to assign a truth value to this statement. Even if the actual die under the cup only ever takes on the values between 1 and 6, so that “26” is not found anywhere under the cup. The statement is still true if and only if the die is showing 3; that is its empirical truth-condition.

---

6. <http://yudkowsky.net/bayes/truthb.html>

7. Page 915, ‘Probability is Subjectively Objective’.

And what about the statement  $((3 * 7) + 5) = 26$ ? Where is the truth-condition for *that* statement located? This I don't know; but I am nonetheless quite confident that it is true. Even though I am not confident that this 'true' means exactly the same thing as the 'true' in "the bucket is 'true' when it contains the same number of pebbles as sheep in the pasture".

So if someone I trust—presumably someone I *really* trust—tells me, "If Oswald hadn't shot Kennedy, someone else would have", and I believe this statement, then I believe the empirical reality is such as to make the counterfactual computation come out this way. Which would seem to imply the conspiracy. And I will anticipate accordingly.

Or if I find out that there *was* a conspiracy, then this will *confirm the truth-condition of the counterfactual*—which might make a bit more sense than saying, "Confirm that the counterfactual is true."

But how do you *actually* compute a counterfactual? For this you must consult Judea Pearl. Roughly speaking, you perform surgery on graphical models of causal processes; you sever some variables from their ordinary parents and surgically set them to new values, and then recalculate the probability distribution.

There are other ways of defining counterfactuals, but I confess they all strike me as entirely odd. Even worse, you have philosophers arguing over what the value of a counterfactual *really is* or *really means*, as if there were some counterfactual world actually floating out there in the philosophical void. If you think I'm attacking a strawperson here, I invite you to consult the philosophical literature on Newcomb's Problem<sup>8</sup>.

A lot of philosophy seems to me to suffer from "naive philosophical realism"—the belief that philosophical debates are about things that automatically and directly exist as propertied objects floating out there in the void.

You can talk about an ideal computation, or an ideal process, that would ideally be applied to the empirical world. You can talk about your uncertain beliefs about the output of this ideal computation, or the result of the ideal process.

So long as the computation is fixed, and so long as the computational itself is only over actually existent things. Or the results of

---

8. [http://lesswrong.com/lw/nc/newcombs\\_problem\\_and\\_regret\\_of\\_rationality/](http://lesswrong.com/lw/nc/newcombs_problem_and_regret_of_rationality/)

other computations previously defined—you should not have your computation be over “nearby possible worlds” unless you can tell me how to compute those, as well.

A chief sign of naive philosophical realism is that it does not tell you how to write a computer program that computes the objects of its discussion.

I have yet to see a camera that peers into “nearby possible worlds”—so even after you’ve analyzed counterfactuals in terms of “nearby possible worlds”, I still can’t write an AI that computes counterfactuals.

But Judea Pearl tells me just how to compute a counterfactual, given only my beliefs about the *actual* world.

I strongly privilege the *real world that actually exists*, and to a slightly lesser degree, logical truths about mathematical objects (preferably finite ones). Anything *else* you want to talk about, I need to figure out how to describe in terms of the first two—for example, as the output of an ideal computation run over the empirical state of the real universe.

The absence of this requirement as a condition, or at least a goal, of modern philosophy, is one of the primary reasons why modern philosophy is often surprisingly useless in my AI work. I’ve read whole books about decision theory that take counterfactual distributions as givens, and never tell you how to compute the counterfactuals.

Oh, and to talk about “the probability that John F. Kennedy was shot, given that Lee Harvey Oswald didn’t shoot him”, we write:

$$P(\text{Kennedy\_shot} \mid \text{Oswald\_not})$$

And to talk about “the probability that John F. Kennedy would have been shot, if Lee Harvey Oswald hadn’t shot him”, we write:

$$P(\text{Oswald\_not} \ [ ] \rightarrow \text{Kennedy\_shot})$$

That little symbol there is supposed to be a box with an arrow coming out of it, but I don’t think Unicode has it.

## 22. Math is Subjunctively Objective<sup>1</sup>

**Followup to:** Probability is Subjectively Objective<sup>2</sup>, Can Counterfactuals Be True?<sup>3</sup>

I am quite confident that the statement  $2 + 3 = 5$  is *true*; I am far less confident of what it *means* for a mathematical statement to be true.

In “The Simple Truth<sup>4</sup>” I defined a pebble-and-bucket system for tracking sheep, and defined a condition for whether a bucket’s pebble level is “true” in terms of the sheep. The bucket is the belief, the sheep are the reality. I believe  $2 + 3 = 5$ . Not just that two sheep plus three sheep equal five sheep, but that  $2 + 3 = 5$ . That is my belief, but where is the reality?

So now the one comes to me and says: “Yes, two sheep plus three sheep equals five sheep, and two stars plus three stars equals five stars. I won’t deny that. But this notion that  $2 + 3 = 5$ , *exists only in your imagination, and is purely subjective.*”

So I say: Excuse me, *what?*

And the one says: “Well, I know what it means to observe two sheep and three sheep leave the fold, and five sheep come back. I know what it means to press ‘2’ and ‘+’ and ‘3’ on a calculator, and see the screen flash ‘5’. I even know what it means to ask someone ‘What is two plus three?’ and hear them say ‘Five.’ But you insist that there is some fact *beyond* this. You insist that  $2 + 3 = 5$ . ”

Well, it kinda *is*.

“Perhaps you just mean that when you *mentally visualize* adding two dots and three dots, you end up visualizing five dots. Perhaps this is the content of what you mean by saying,  $2 + 3 = 5$ . I have no trouble with that, for brains are as real as sheep.”

No, for it seems to me that  $2 + 3$  equaled  $5$  *before* there were any humans around to do addition. When humans showed up on the scene, they did not *make*  $2 + 3$  equal  $5$  by virtue of thinking it. Rather, they thought that ‘ $2 + 3 = 5$ ’ *because*  $2 + 3$  did in fact equal  $5$ .

“Prove it.”

---

1. [http://lesswrong.com/lw/si/math\\_is\\_subjunctively\\_objective/](http://lesswrong.com/lw/si/math_is_subjunctively_objective/)

2. Page 915, ‘Probability is Subjectively Objective’.

3. Page 944, ‘Can Counterfactuals Be True?’.

4. <http://yudkowsky.net/bayes/truth.html>

I'd love to, but I'm busy; I've got to, um, eat a salad.

"The *reason you believe* that  $2 + 3 = 5$ , is your mental visualization of two dots plus three dots yielding five dots. Does this not imply that this physical event in your physical brain is the *meaning* of the statement ' $2 + 3 = 5$ '?"

But I honestly don't think that *is* what I mean. Suppose that by an amazing cosmic coincidence, a flurry of neutrinos struck my neurons, causing me to imagine two dots colliding with three dots and visualize six dots. I would then say, ' $2 + 3 = 6$ '. But this wouldn't mean that  $2 + 3$  actually *had* become equal to 6. Now, if what I mean by ' $2 + 3$ ' consists entirely of what my mere physical brain merely *happens to output*, then a neutrino *could* make  $2 + 3 = 6$ . But you can't change arithmetic by tampering with a calculator.

"Aha! I have you now!"

Is that so?

"Yes, you've given your whole game away!"

Do tell.

"You visualize a subjunctive world, a counterfactual<sup>5</sup>, where your brain is struck by neutrinos, and says, ' $2 + 3 = 6$ '. So you know that in this case, your future self will *say* that ' $2 + 3 = 6$ '. But then you add up dots in your *own, current brain*, and your *current* self gets five dots. So you say: 'Even if I believed " $2 + 3 = 6$ ", then  $2 + 3$  would still equal 5.' You say: ' $2 + 3 = 5$  regardless of what anyone thinks of it.' So your *current* brain, computing the same question while it *imagines* being different but is not *actually* different, finds that the answer *seems to be the same*. Thus your brain creates the *illusion* of an additional reality that exists outside it, independent of any brain."

Now hold on! You've *explained* my belief that  $2 + 3 = 5$  regardless of what anyone thinks, but that's not the same as explaining away<sup>6</sup> my belief. Since  $2 + 3 = 5$  does not, *in fact*, depend on what any human being thinks of it, therefore it is *right and proper* that when I imagine counterfactual<sup>7</sup> worlds in which people (including myself) *think* ' $2 + 3 = 6$ ', and I ask what  $2 + 3$  *actually* equals in this counterfactual world, it still comes out as 5.

5. Page 944, 'Can Counterfactuals Be True?'.

6. Page 703, 'Explaining vs. Explaining Away'.

7. Page 944, 'Can Counterfactuals Be True?'.

“Don’t you see, that’s just like trying to visualize motion stopping everywhere in the universe, by imagining yourself as an observer outside the universe who experiences time passing while nothing moves<sup>8</sup>. But really there is no time without motion.”

I see the analogy, but I’m not sure it’s a deep analogy<sup>9</sup>. Not everything you can imagine seeing, doesn’t exist. It seems to me that a brain can *easily* compute quantities that don’t depend on the brain.

“*What?* Of course everything that the brain computes depends on the brain! Everything that the brain computes, is computed inside the brain!”

That’s not what I mean! I just mean that the brain can perform computations that *refer to* quantities outside the brain. You can set up a question, like ‘How many sheep are in the field?’, that isn’t *about* any particular person’s brain, and whose *actual* answer doesn’t *depend on* any particular person’s brain. And then a brain can faithfully compute that answer.

If I count two sheep and three sheep returning from the field, and Autrey<sup>10</sup>’s brain gets hit by neutrinos so that Autrey thinks there are six sheep in the fold, then that’s not going to *cause* there to be six sheep in the fold—right? The whole question here is just *not about* what Autrey thinks, it’s *about* how many sheep are in the fold.

Why should I care what *my* subjunctive future self thinks is the sum of  $2 + 3$ , any more than I care what *Autrey* thinks is the sum of  $2 + 3$ , when it comes to asking what is *really* the sum of  $2 + 3$ ?

“Okay... I’ll take another tack. Suppose you’re a psychiatrist, right? And you’re an expert witness in court cases—basically a hired gun, but you try to deceive yourself about it. Now wouldn’t it be a bit suspicious<sup>11</sup>, to find yourself saying: ‘Well, the only reason *that I in fact believe* that the defendant is insane, is because I was paid to be an expert psychiatric witness for the defense. And if I had been paid to witness for the prosecution, I undoubtedly would have come to the conclusion that the defendant is sane. But my belief that the defendant is insane, is *perfectly justified*; it is justified by

---

8. [http://lesswrong.com/lw/qp/timeless\\_physics/](http://lesswrong.com/lw/qp/timeless_physics/)

9. [http://lesswrong.com/lw/rj/surface\\_analogies\\_and\\_deep\\_causes/](http://lesswrong.com/lw/rj/surface_analogies_and_deep_causes/)

10. <http://yudkowsky.net/bayes/truth.html>

11. Page 513, ‘The Bottom Line’.

my observation that the defendant used his own blood to paint an Elder Sign on the wall of his jail cell.”

Yes, that *does* sound suspicious, but I don’t see the point.

“My point is that the *physical cause* of your belief that  $2 + 3 = 5$ , is the physical event of your brain visualizing two dots and three dots and coming up with five dots. If your brain came up six dots, due to a neutrino storm or whatever, you’d think ‘ $2 + 3 = 6$ ’. How can you possibly say that your belief *means* anything other than the number of dots your brain came up with?”

Now hold on just a second. Let’s say that the psychiatrist is paid by the judge, and when he’s paid by the judge, he renders an honest and neutral evaluation, and his evaluation is that the defendant is sane, just played a bit too much Mythos. So it is true to say that if<sup>12</sup> the psychiatrist had been paid by the defense, then the psychiatrist would have found the defendant to be insane. But that doesn’t mean that when the psychiatrist is paid by the *judge*, you should dismiss his evaluation as telling you *nothing more than* ‘the psychiatrist was paid by the judge’. On those occasions where the psychiatrist *is* paid by the judge, his opinion varies with the defendant, and conveys real evidence about the defendant.

“Okay, so now what’s *your* point?”

That when my brain is *not* being hit by a neutrino storm, it yields honest and informative evidence that  $2 + 3 = 5$ .

“And if your brain *was* hit by a neutrino storm, you’d be saying, ‘ $2 + 3 = 6$  regardless of what anyone thinks of it’. Which shows how reliable *that* line of reasoning is.”

I’m not claiming that my saying ‘ $2 + 3 = 5$  no matter what anyone thinks’ represents stronger *numerical* evidence than my saying ‘ $2 + 3 = 5$ ’. My saying the former just tells you something extra about my epistemology, not numbers.

“And you don’t think your epistemology is, oh, a little... *incoherent*? ”

No! I think it is perfectly coherent to simultaneously hold all of the following:

- $2 + 3 = 5$ .
- If neutrinos make me believe “ $2 + 3 = 6$ ”, then  $2 + 3 = 5$ .

---

<sup>12</sup>. Page 944, ‘Can Counterfactuals Be True?’.

- If neutrinos make me believe “ $2 + 3 = 6$ ”, then I will say “ $2 + 3 = 6$ ”.
- If neutrinos make me believe that “ $2 + 3 = 6$ ”, then I will thereafter assert that “If neutrinos make me believe ‘ $2 + 3 = 5$ ’, then  $2 + 3 = 6$ ”.
- The cause of my thinking that “ $2 + 3 = 5$  independently of what anyone thinks” is that my *current* mind, when it subjunctively recomputes the value of  $2 + 3$  under the assumption that my *imagined* self is hit by neutrinos, does not see the *imagined* self’s beliefs as changing the dots, and my *current* brain just visualizes two dots plus three dots, as before, so that the imagination of my *current* brain shows the same result.
- If I were *actually* hit by neutrinos, my brain would compute a different result, and I would assert “ $2 + 3 = 6$  independently of what anyone thinks.”
- $2 + 3 = 5$  independently of what anyone thinks.
- Since  $2 + 3$  will *in fact* go on equaling 5 *regardless* of what I imagine about it or how my brain visualizes cases where my future self has different beliefs, it’s a *good thing* that my imagination doesn’t visualize the result as depending on my beliefs.

“Now that’s just crazy talk!”

No, you’re the crazy one! You’re *collapsing your levels*; you think that just because my brain asks a question, it should start mixing up queries about the state of my brain *into* the question. Not every question my brain asks is *about* my brain!

Just because something is computed *in* my brain, doesn’t mean that my computation has to depend on my brain’s *representation of* my brain. It certainly doesn’t mean that the *actual quantity* depends on my brain! It’s my brain that computes my beliefs about gravity, and if neutrinos hit me I will come to a different conclusion; but that doesn’t mean that I can think different and fly. And I don’t *think* I can think different and fly, either!

I am not a calculator who, when someone presses my “2” and “+” and “3” buttons, computes, “What do I output when someone presses  $2 + 3$ ?” I am a calculator who computes “What is  $2 + 3$ ?” The former is a circular question that can consistently return any answer<sup>13</sup>—which makes it not very *helpful*.

Shouldn't we expect non-circular questions to be the *normal* case? The brain evolved to guess at the state of the environment, not guess at 'what the brain will think is the state of the environment'. Even when the brain models itself, it is trying to *know itself*, not trying to know *what it will think about itself*.

Judgments that depend on our representations of *anyone's* state of mind, like "It's okay to kiss someone only if they want to be kissed", are the exception rather than the rule.

*Most* quantities we bother to think about at all, will appear to be 'the same regardless of what anyone thinks of them'. When we imagine thinking differently about the quantity, we will imagine the quantity coming out the same; it will feel "subjunctively objective".

And there's nothing wrong with that! If something *appears* to be the same regardless of what anyone thinks, then maybe that's because it *actually is* the same regardless of what anyone thinks.

Even if you explain that the quantity *appears* to stay the same in my imagination, *merely* because my current brain computes it the same way—well, how *else* would I imagine something, *except* with my current brain? Should I imagine it using a rock?

"Okay, so it's possible for something that appears thought-independent, to actually be thought-independent. But why do you think that  $2 + 3 = 5$ , in particular, has some kind of existence independently of the dots you imagine?"

Because two sheep plus three sheep equals five sheep, and this appears to be true in every mountain and every island, every swamp and every plain and every forest.

And moreover, it is also true of two rocks plus three rocks.

And further, when I press buttons upon a calculator and activate a network of transistors, it *successfully predicts* how many sheep or rocks I will find.

Since all these quantities, correlate with each other and successfully predict each other, surely they must have something *like* a common cause, a similarity that factors out? Something that is true beyond and before the concrete observations? Something that the concrete observations hold in common? And this commonality is then also the sponsor of my answer, 'five', that I find in my own brain.

---

13. Page 915, 'Probability is Subjectively Objective'.

“But my dear sir, if the fact of  $2 + 3 = 5$  exists somewhere outside your brain... *then where is it?*”

Damned if I know.

## 23. Does Your Morality Care What You Think?<sup>1</sup>

**Followup to:** Math is Subjunctively Objective<sup>2</sup>, The Moral Void<sup>3</sup>, Is Morality Given<sup>4</sup>

Thus I recall the study, though I cannot recall the citation:

Children, at some relatively young age, were found to distinguish between:

- The teacher, by saying that we're allowed to stand on our desks, can make it right to do so.
- The teacher, by saying that I'm allowed to take something from another child's backpack, *cannot* make it right to do so.

Obert: "Well, I don't know the citation, but it sounds like a fascinating study. So even children, then, realize that moral facts are givens<sup>5</sup>, beyond the ability of teachers or parents to alter."

Subhan: "You say that like it's a good thing. Children may also think that people in Australia have to wear heavy boots from falling off the other side of the Earth."

Obert: "Call me Peter Pan, then, because I never grew up on this one. Of course it doesn't matter what the teacher says. It doesn't matter what I say. It doesn't even matter what I *think*. Stealing is wrong. Do you *disagree*?"

Subhan: "You don't see me picking your pockets, do you? Isn't it enough that I *choose* not to steal from you—do I have to pretend it's the law of the universe?"

Obert: "Yes, or I can't trust your commitment."

Subhan: "A... revealing remark. But really, I don't think that this experimental result seems at all confusing, in light of the recent

---

1. [http://lesswrong.com/lw/sj/does\\_your\\_morality\\_care\\_what\\_you\\_think/](http://lesswrong.com/lw/sj/does_your_morality_care_what_you_think/)

2. Page 948, 'Math is Subjunctively Objective'.

3. Page 856, 'The Moral Void'.

4. Page 879, 'Is Morality Given?'.

5. Page 879, 'Is Morality Given?'.

discussion of subjunctive objectivity<sup>6</sup>—a discussion in which Eliezer strongly supported my position, by the way.”

Obert: “Really? I thought Eliezer was finally coming out in favor of *my* position.”

Subhan: “Huh? How do you get *that*? ”

Obert: “The whole subtext of ‘Math is Subjunctively Objective’<sup>7</sup> is that morality is just like math! Sure, we compute morality inside our own brains—where else would we compute it? But just because we compute a quantity inside our own brains, doesn’t mean that *what is computed* has a dependency on our own state of mind.”

Subhan: “I think we must have been reading different *Overcoming Bias* posts! The whole subtext of ‘Math is Subjunctively Objective’<sup>8</sup> is to explain *away* why morality *seems* objective—to show that the *feeling* of a fixed given can arise without any external referent. When you *imagine* yourself thinking that killing is right, your brain—that imagines hasn’t yet been altered, so you carry out that moral imagination with your *current* brain, and conclude: ‘Even if I thought killing were right, killing would still be wrong.’ But this *doesn’t* show that killing-is-wrong is a fixed fact from outside you.”

Obert: “Like, say,  $2 + 3 = 5$  is a fixed fact. Eliezer wrote: ‘If something *appears* to be the same regardless of what anyone thinks, then maybe that’s because it *actually is* the same regardless of what anyone thinks.’ I’d say that subtext is pretty clear!”

Subhan: “On the contrary. Naively, you might imagine your future self thinking differently of a thing, and visualize that the thing wouldn’t thereby change, and conclude that the thing existed outside you. Eliezer shows how this is not *necessarily* the case. So you shouldn’t *trust your intuition* that the thing is objective—it might be that the thing exists outside you, or it might *not*. It has to be *argued separately* from the feeling of subjunctive objectivity. In the case of  $2 + 3 = 5$ , it’s at least reasonable to wonder if math existed before humans. Physics itself seems to be made of math, and if we don’t tell a story where physics was around before humans could observe it, it’s hard to give a coherent account of how we got here. But there’s not the slightest evidence that *morality* was at work in the universe before humans got here. *We created it.*”

---

6. Page 948, ‘Math is Subjunctively Objective’.

7. Page 948, ‘Math is Subjunctively Objective’.

8. Page 948, ‘Math is Subjunctively Objective’.

Obert: “I know some very wise children who would disagree with you.”

Subhan: “Then they’re wrong! If children learned in school that it was okay to steal, they would grow up believing it was okay to steal.”

Obert: “Not if they saw that stealing hurt the other person, and felt empathy for their pain. Empathy is a human universal<sup>9</sup>. ”

Subhan: “So we take a step back and say that evolution created the emotions that gave rise to morality<sup>10</sup>, it doesn’t put morality anywhere outside us. But what you say might not even be true—if theft weren’t considered a crime, the other child might not feel so hurt by it. And regardless, it is rare to find any child capable of fully reconsidering the moral teachings of its society.”

Obert: “I hear that, in a *remarkable similarity* to Eliezer, your parents were Orthodox Jewish and you broke with religion as a very young child.”

Subhan: “I doubt that I was internally generating *de novo* moral philosophy. I was probably just wielding, against Judaism, the morality of the science fiction that actually socialized me.”

Obert: “Perhaps you underestimate yourself. How much science fiction had you read at the age of five, when you realized it was dumb to recite Hebrew prayers you couldn’t understand? Children may see errors that adults are too adept at fooling themselves to realize<sup>11</sup>. ”

Subhan: “Hah! In all probability, if the teacher *had in fact* said that it was okay to take things from other children’s backpacks, the children *would in fact* have thought it was right to steal.”

Obert: “Even if true, that doesn’t prove anything. It is quite coherent to simultaneously hold that:<sup>12</sup>”

- “Stealing is wrong.”
- “If a neutrino storm makes me believe ‘stealing is right’, then stealing is wrong.”

---

9. [http://lesswrong.com/lw/r/the\\_psychological\\_unity\\_of\\_humankind/](http://lesswrong.com/lw/r/the_psychological_unity_of_humankind/)

10. Page 927, ‘The Gift We Give To Tomorrow’.

11. Page 505, ‘Knowing About Biases Can Hurt People’.

12. Page 948, ‘Math is Subjunctively Objective’.

- “If a neutrino storm makes me believe ‘stealing is right’, then I will say, ‘If a neutrino storm makes me believe “stealing is wrong”, then stealing is right.’”

Subhan: “Fine, it’s *coherent*, but that doesn’t mean it’s *true*. The morality that the child has *in fact* learned from the teacher—or their parents, or the other children, or the television, or their parents’ science fiction collection—doesn’t say, ‘Don’t steal *because the teacher says so*.’ The learned morality just says, ‘Don’t steal.’ The cognitive procedure by which the children were taught to judge, does not have an internal dependency on what the children believe the teacher believes. That’s why, in their moral imagination, it feels objective. But where did they acquire that morality in the first place? From the teacher!”

Obert: “So? I don’t understand—you’re saying that because they learned about morality from the teacher, they should think that morality has to be *about* the teacher? That they should think the teacher has the power to make it right to steal? How does that follow? It is quite coherent to simultaneously hold that—”

Subhan: “I’m saying that they got the morality *from the teacher!* Not from some mysterious light in the sky!”

Obert: “Look, I too read science fiction and fantasy as a child, and I think I may have been to some degree socialized by it—”

Subhan: “What a *remarkable coincidence*.”

Obert: “The stories taught me that it was right to care about people who were different from me—aliens with strange shapes, aliens made of something other than carbon atoms, AIs who had been created rather than evolved, even things that didn’t think like a human. But none of the stories ever said, ‘You should care about people of different shapes and substrates *because science fiction told you to do it, and what science fiction says, goes*.’ I wouldn’t have bought that.”

Subhan: “Are you sure you wouldn’t have? That’s how religion works.”

Obert: “Didn’t work on you. Anyway, the novels said to care about the aliens *because* they had inner lives and joys—or *because* I wouldn’t want aliens to mistreat humans—or *because* shape and substrate never had anything to do with what makes a person a person. And you know, that still seems to me like a good justification.”

Subhan: “Of course; you were *told* it was a good justification—maybe not directly, but the author showed other characters responding to the argument.”

Obert: “It’s not like the science fiction writers were making up their morality from scratch. They were working at the end of a chain of moral arguments and debates that stretches back to the Greeks, probably to before writing, maybe to before the dawn of modern humanity. You can *learn* morality, not just get pressed into it like a Jello mold. If you learn  $2 + 3 = 5$  from a teacher, it doesn’t mean the teacher has the power to add two sheep to three sheep and get six sheep. If you would have spouted back ‘ $2 + 3 = 6$ ’ if the teacher said so, that doesn’t change the sheep, it just means that you don’t really understand the subject<sup>13</sup>. So too with morality.”

Subhan: “Okay, let me try a different tack. You, I take it, agree with both of these statements:”

- “If I preferred to kill people, it would not become right to kill people.”
- “If I preferred to eat anchovy pizzas, it would become right to eat anchovy pizzas.”

Obert: “Well, there are various caveats I’d attach to both of those. Like, in any circumstance where I really did prefer to kill someone, there’d be a high probability he was about to shoot me, or something. And there’s all kinds of ways that eating an anchovy pizza could be wrong, like if I was already overweight. And I don’t claim to be certain of anything when it comes to morality<sup>14</sup>. But on the whole, and omitting all objections and knock-on effects, I agree.”

Subhan: “It’s that second statement I’m really interested in. How does your wanting to eat an anchovy pizza *make* it right?”

Obert: “Because *ceteris paribus*, in the course of ordinary life as we know it, and barring unspecified side effects, it is good for sentient beings to get what they want.”

Subhan: “And why doesn’t that apply to the bit about killing, then?”

Obert: “Because the other person doesn’t want to die. Look, the whole reason why it’s right *in the first place* for me to eat pepper-

<sup>13</sup>. Page 140, ‘Guessing the Teacher’s Password’.

<sup>14</sup>. Page 935, ‘Could Anything Be Right?’.

oni pizza—the *original justification*—is that I enjoy doing so. Eating pepperoni pizza makes me happy, which is *ceteris paribus* a good thing. And eating anchovy pizza—blegh! *Ceteris paribus*, it's not good for sentient beings to experience disgusting tastes. But if my taste in pizza changes, that changes the consequences of eating, which changes the moral justification, and so the moral judgment changes as well. But the reasons for not *killing* are in terms of the other person having an inner life that gets snuffed out—a fact that doesn't change depending on my own state of mind.”

Subhan: “Oh? I was guessing that the difference had something to do with the social disapproval that would be leveled at murder, but not at eating anchovy pizza.”

Obert: “As usual, your awkward attempts at rationalism have put you out of touch with self-evident moral truths. That’s just not how I, or other real people, actually think! If I want to *bleep bleep bleep* a consenting adult, it doesn’t matter whether society approves. Society can go *bleep bleep bleep bleep bleep* -”

Subhan: “Or so science fiction taught you.”

Obert: “Spider Robinson’s science fiction, to be precise. ‘Whatever turns you on’ shall be the whole of the law. So long as the ‘you’ is plural.”

Subhan: “So that’s where you got that particular self-evident moral truth. Was it also Spider Robinson who told you that it was self-evident?”

Obert: “No, I thought about that for a while, and then decided myself.”

Subhan: “You seem to be paying remarkably close attention to what people *want*. Yet you insist that what validates this attention, is some external standard that makes the satisfaction of desires, *good*. Can’t you just admit that, by empathy and vicarious experience and evolved fellow-feeling, you want others to get what they want? When does this external standard ever say that it’s good for something to happen that someone *doesn’t want*? ”

Obert: “Every time you’ve got to tell your child to lay off the ice cream, he’ll grow more fat cells that will make it impossible for him to lose weight as an adult.”

Subhan: “And could something good happen that *no one* wanted?”

Obert: “I rather expect so. I don’t think we’re all *entirely* past our childhoods. In some ways the human species itself strikes me as being a sort of toddler in the ‘No!’ stage.”

Subhan: “Look, there’s a perfectly normal and non-mysterious chain of causality that describes where morality comes from, and it’s not from outside humans<sup>15</sup>. If you’d been told that killing was right, or if you’d evolved to enjoy killing—much more than we already do, I mean—or if you *really did* have a mini-stroke that damaged your frontal lobe, then you’d be going around saying, ‘Killing is right regardless of what anyone thinks of it’. No great light in the sky would correct you. There is nothing else to the story.”

Obert: “Really, I think that in this whole debate between us, there is surprisingly little information to be gained by such observations as ‘You only say that because your brain makes you say it.’<sup>16</sup> If a neutrino storm hit me, I might say ‘ $2 + 3 = 6$ ’, but that wouldn’t change arithmetic. It would just make my brain compute something other than arithmetic. And these various misfortunes that you’ve described, wouldn’t change the crime of murder. They would just make my brain compute something other than morality.”

---

15. Page 907, ‘Rebelling Within Nature’.

16. [http://lesswrong.com/lw/ro/thou\\_art\\_physics/](http://lesswrong.com/lw/ro/thou_art_physics/)

## 24. Changing Your Metaethics<sup>1</sup>

**Followup to:** The Moral Void<sup>2</sup>, Joy in the Merely Real<sup>3</sup>, No Universally Compelling Arguments<sup>4</sup>, Where Recursive Justification Hits Bottom<sup>5</sup>, The Gift We Give To Tomorrow<sup>6</sup>, Does Your Morality Care What You Think?<sup>7</sup>, Existential Angst Factory<sup>8</sup>, ...

If you say, “Killing people is wrong,” that’s morality. If you say, “You shouldn’t kill people because God prohibited it,” or “You shouldn’t kill people because it goes against the trend of the universe”, that’s metaethics.

Just as there’s far more agreement on Special Relativity than there is on the question “What is science?<sup>9</sup>”, people find it much easier to agree “Murder is bad” than to agree *what* makes it bad, or what it *means* for something to be bad.

People do get attached<sup>10</sup> to their metaethics. Indeed they frequently insist that if their metaethic is wrong, all morality necessarily falls apart. It might be interesting to set up a panel of metaethicists—theists, Objectivists, Platonists, etc.—all of whom agree that killing is wrong; all of whom disagree on what it means for a thing to be “wrong”; and all of whom insist that if their metaethic is untrue, then morality falls apart.

Clearly a good number of people, if they are to make philosophical progress, will need to shift metathics at some point in their lives. *You* may have to do it.

At that point, it might be useful to have an open line of retreat<sup>11</sup>—not a retreat from morality, but a retreat from Your-Current-Metaethic. (You know, the one that, if it is not true, leaves no possible basis for not killing people.)

---

1. [http://lesswrong.com/lw/sk/changing\\_your\\_metaethics/](http://lesswrong.com/lw/sk/changing_your_metaethics/)

2. Page 856, ‘The Moral Void’.

3. Page 715, ‘Joy in the Merely Real’.

4. Page 844, ‘No Universally Compelling Arguments’.

5. Page 887, ‘Where Recursive Justification Hits Bottom’.

6. Page 927, ‘The Gift We Give To Tomorrow’.

7. Page 955, ‘Does Your Morality Care What You Think?’.

8. Page 940, ‘Existential Angst Factory’.

9. [http://lesswrong.com/lw/in/scientific\\_evidence\\_legal\\_evidence\\_rational/](http://lesswrong.com/lw/in/scientific_evidence_legal_evidence_rational/)

10. Page 384, ‘Affective Death Spirals’.

11. Page 653, ‘Leave a Line of Retreat’.

And so I've been setting up these lines of retreat, in many and various posts, summarized below. For I have learned that to change metaethical beliefs is nigh-impossible in the presence of an unanswered attachment.

If, for example, someone believes the authority of "Thou Shalt Not Kill" derives from God, then there are several and well-known things to say that can help set up a line of retreat—as opposed to immediately attacking the plausibility of God. You can say, "Take personal responsibility!<sup>12</sup> Even if you got orders from God, it would be your own decision to obey those orders. Even if God didn't order you to be moral, you could just be moral anyway."

The above argument actually generalizes to quite a number of metaethics—you just substitute Their-Favorite-Source-Of-Morality, or even the word "morality", for "God". Even if your particular source of moral authority failed, couldn't you just drag the child off the train tracks *anyway*? And indeed, who is it but you, that ever decided to follow this source of moral authority in the first place? What responsibility are you really passing on?

So the most important line of retreat is the one given in *The Moral Void*<sup>13</sup>: If your metaethic stops telling you to save lives, you can just drag the kid off the train tracks anyway. To paraphrase Piers Anthony, only those who have moralities worry over whether or not they have them.<sup>14</sup> If your metaethic tells you to kill people, why *should* you even listen? Maybe that which you would do even if there were no morality, *is* your morality.

The point being, of course, not that no morality exists; but that you can hold your will in place, and not fear losing sight of what's important to you<sup>15</sup>, while your notions of the *nature* of morality change.

Other posts are there to set up lines of retreat specifically for more *naturalistic* metaethics. It may make more sense where I'm coming from on these, once I *actually* present my metaethic; but I thought it wiser to set them up in advance, to leave lines of retreat.

---

12. Page 879, 'Is Morality Given?'.

13. Page 856, 'The Moral Void'.

14. [http://lesswrong.com/lw/ky/fake\\_morality/](http://lesswrong.com/lw/ky/fake_morality/)

15. Page 644, 'Something to Protect'.

Joy in the Merely Real<sup>16</sup> and Explaining vs. Explaining Away<sup>17</sup> argue that you shouldn't be disappointed in any facet of life, just because it turns out to be *explicable* instead of inherently mysterious<sup>18</sup>: for if we cannot take joy in the merely real, our lives shall be empty indeed.

No Universally Compelling Arguments<sup>19</sup> sets up a line of retreat from the desire to have *everyone* agree with our moral arguments. There's a strong moral intuition which says that if our moral arguments are right, by golly, we ought to be able to *explain* them to people. This may be valid among humans<sup>20</sup>, but you can't explain moral arguments to a rock. There is no ideal philosophy student of perfect emptiness who can be persuaded to implement modus ponens, starting without modus ponens<sup>21</sup>. If a mind doesn't contain that which is moved by your moral arguments, it won't respond to them.

But then isn't all morality circular logic, in which case it falls apart? Where Recursive Justification Hits Bottom<sup>22</sup> and My Kind of Reflection<sup>23</sup> explain the difference between a self-consistent loop through the meta-level, and actual circular logic. You shouldn't find yourself saying "The universe is simple because it is simple", or "Murder is wrong because it is wrong"; but neither should you try to abandon Occam's Razor while evaluating the probability that Occam's Razor works, nor should you try to evaluate "Is murder wrong?" from somewhere outside your brain. There is no ideal philosophy student of perfect emptiness to which you can unwind yourself—try to find the perfect rock to stand upon, and you'll end up as a rock. So instead use the full force of your intelligence, your full rationality and your full morality, when you investigate the foundations of yourself.

The Gift We Give To Tomorrow<sup>24</sup> sets up a line of retreat for those afraid to allow a *causal* role for evolution, in their account of

16. Page 715, 'Joy in the Merely Real'.

17. Page 703, 'Explaining vs. Explaining Away'.

18. Page 154, 'Mysterious Answers to Mysterious Questions'.

19. Page 844, 'No Universally Compelling Arguments'.

20. [http://lesswrong.com/lw/r/the\\_psychological\\_unity\\_of\\_humankind/](http://lesswrong.com/lw/r/the_psychological_unity_of_humankind/)

21. Page 861, 'Created Already In Motion'.

22. Page 887, 'Where Recursive Justification Hits Bottom'.

23. Page 896, 'My Kind of Reflection'.

24. Page 927, 'The Gift We Give To Tomorrow'.

how morality came to be. (Note that this is extremely distinct from granting evolution a *justificational* status in moral theories.) Love has to come into existence somehow—for if we cannot take joy in things that can come into existence, our lives will be empty indeed. Evolution may not be a particularly *pleasant* way for love to evolve, but judge the end product—not the source. Otherwise you would be committing what is known (appropriately) as The Genetic Fallacy<sup>25</sup>: causation is not the same concept as justification. It's not like you can step outside the brain evolution gave you: Rebelling against nature is only possible from within nature.<sup>26</sup>

The earlier series on Evolutionary Psychology<sup>27</sup> should dispense with the metaethical confusion of believing that any normal human being thinks about their reproductive fitness, even unconsciously, in the course of making decisions. Only evolutionary biologists even know how to *define* genetic fitness, and they know better than to think it defines morality.

Alarming indeed is the thought that morality might be computed inside our own minds—doesn't this imply that morality is a mere thought? Doesn't it imply that whatever you think is right, must be right? Posts such as Does Your Morality Care What You Think?<sup>28</sup> and its predecessors, Math is Subjunctively Objective<sup>29</sup> and Probability is Subjectively Objective<sup>30</sup>, set up the needed line of retreat: Just because a quantity is computed inside your head, doesn't mean that the quantity computed is *about* your thoughts. There's a difference between a calculator that calculates “What is  $2 + 3$ ?” and “What do I output when someone presses ‘2’, ‘+’, and ‘3’?”

And finally Existential Angst Factory<sup>31</sup> offers the notion that if life seems painful, reductionism may not be the real source of your problem—if living in a world of mere particles seems too unbearable, maybe your life isn't exciting enough on its own?

If all goes well, my next post will set up the metaethical question and its methodology, and I'll present my actual answer on Monday.

25. Page 485, ‘The Genetic Fallacy’.

26. Page 907, ‘Rebelling Within Nature’.

27. [http://lesswrong.com/lw/l1/evolutionary\\_psychology/](http://lesswrong.com/lw/l1/evolutionary_psychology/)

28. Page 955, ‘Does Your Morality Care What You Think?’.

29. Page 948, ‘Math is Subjunctively Objective’.

30. Page 915, ‘Probability is Subjectively Objective’.

31. Page 940, ‘Existential Angst Factory’.

And if you're wondering why I deem this business of metaethics important, when it is all going to end up adding up to moral normality<sup>32</sup>... telling you to pull the child off the train tracks, rather than the converse<sup>33</sup>...

Well, there *is* opposition to rationality from people who think it drains meaning from the universe.

And this is a special case of a general phenomenon, in which many many people get messed up by misunderstanding where their morality comes from. Poor metaethics forms part of the teachings of many a cult<sup>34</sup>, including the big ones<sup>35</sup>. My target audience is not just people who are afraid that life is meaningless, but also those who've concluded that love is a delusion because real morality has to involve maximizing your inclusive fitness, or those who've concluded that unreturned kindness is evil because real morality arises only from selfishness<sup>36</sup>, etc.

But the *real* reason, of course...

---

32. [http://lesswrong.com/lw/qz/living\\_in\\_many\\_worlds/](http://lesswrong.com/lw/qz/living_in_many_worlds/)

33. <http://www.singinst.org/blog/2007/06/16/transhumanism-as-simplified-humanism/>

34. Page 406, 'Every Cause Wants To Be A Cult'.

35. [http://lesswrong.com/lw/i8/religions\\_claim\\_to\\_be\\_nondisprovable/](http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/)

36. [http://lesswrong.com/lw/kx/fake\\_selfishness/](http://lesswrong.com/lw/kx/fake_selfishness/)

## 25. Setting Up Metaethics<sup>1</sup>

**Followup to:** Is Morality Given?<sup>2</sup>, Is Morality Preference?<sup>3</sup>, Moral Complexities<sup>4</sup>, Could Anything Be Right?<sup>5</sup>, The Bedrock of Fairness<sup>6</sup>, ...

Intuitions about morality seem to split up into two broad camps: morality-as-given<sup>7</sup> and morality-as-preference<sup>8</sup>.

Some perceive morality as a *fixed given*, independent of our whims, about which we form changeable *beliefs*. This view's great advantage is that it seems more normal<sup>9</sup> up at the level of everyday moral conversations: it is the intuition underlying our everyday notions of “moral error”, “moral progress”, “moral argument”, or “just because you want to murder someone doesn't make it *right*“.

Others choose to describe morality as a *preference*—as a desire in some particular person; nowhere else is it written<sup>10</sup>. This view's great advantage is that it has an easier time living with reductionism<sup>11</sup>—fitting the notion of “morality” into a universe of mere physics<sup>12</sup>. It has an easier time at the *meta* level, answering questions like “What is morality?” and “Where does morality come from?<sup>13</sup>”

Both intuitions must contend with seemingly impossible questions<sup>14</sup>. For example, Moore's Open Question<sup>15</sup>: Even if you come up with some simple answer that fits on T-Shirt, like “Happiness<sup>16</sup> is the sum total of goodness<sup>17</sup>!”, you would need to *argue* the identi-

---

1. [http://lesswrong.com/lw/sl/setting\\_up\\_metaethics/](http://lesswrong.com/lw/sl/setting_up_metaethics/)

2. Page 879, ‘Is Morality Given?’.

3. Page 871, ‘Is Morality Preference?’.

4. Page 869, ‘Moral Complexities’.

5. Page 935, ‘Could Anything Be Right?’.

6. Page 864, ‘The Bedrock of Fairness’.

7. Page 879, ‘Is Morality Given?’.

8. Page 871, ‘Is Morality Preference?’.

9. [http://lesswrong.com/lw/qz/living\\_in\\_many\\_worlds/](http://lesswrong.com/lw/qz/living_in_many_worlds/)

10. Page 927, ‘The Gift We Give To Tomorrow’.

11. Page 699, ‘Reductionism’.

12. Page 715, ‘Joy in the Merely Real’.

13. Page 927, ‘The Gift We Give To Tomorrow’.

14. Page 679, ‘Wrong Questions’.

15. <http://plato.stanford.edu/entries/metaethics/#IsOOpeQueArg>

16. [http://lesswrong.com/lw/lb/not\\_for\\_the\\_sake\\_of\\_happiness\\_alone/](http://lesswrong.com/lw/lb/not_for_the_sake_of_happiness_alone/)

17. [http://lesswrong.com/lw/lq/fake\\_utility\\_functions/](http://lesswrong.com/lw/lq/fake_utility_functions/)

ty. It isn't instantly obvious to everyone that goodness is happiness, which seems to indicate that happiness and rightness were different concepts to start with. What was that second concept, then, originally?

Or if "Morality is mere preference!" then *why care* about human preferences? How is it possible to establish any "ought" at all, in a universe seemingly of mere "is"?

So what we should want, ideally, is a metaethic that:

1. Adds up to moral normality, including moral errors, moral progress<sup>18</sup>, and things you should do whether you want to or not<sup>19</sup>;
2. Fits naturally into a non-mysterious<sup>20</sup> universe, postulating no exception to reductionism;
3. Does not oversimplify<sup>21</sup> humanity's complicated moral arguments and many terminal values<sup>22</sup>;
4. Answers<sup>23</sup> all the impossible questions.

I'll present that view tomorrow.

Today's post is devoted to setting up the question.

Consider "free will", already<sup>24</sup> dealt<sup>25</sup> with<sup>26</sup> in these posts. On one level of organization, we have mere<sup>27</sup> physics, particles that make no choices. On another level of organization, we have human minds that extrapolate possible futures and choose between them. How can we control anything, even our own choices, when the universe is deterministic?<sup>28</sup>

To dissolve the puzzle of free will, you have to simultaneously imagine two levels of organization while keeping them conceptually distinct. To get it on a gut level, you have to *see* the level transition<sup>29</sup>—the way in which free will *is* how the human decision

18. Page 924, 'Whither Moral Progress?'.

19. Page 955, 'Does Your Morality Care What You Think?'.

20. Page 154, 'Mysterious Answers to Mysterious Questions'.

21. Page 839, 'Heading Toward Morality'.

22. [http://lesswrong.com/lw/l3/thou\\_art\\_godsbatter/](http://lesswrong.com/lw/l3/thou_art_godsbatter/)

23. Page 673, 'Dissolving the Question'.

24. [http://lesswrong.com/lw/r1/timeless\\_control/](http://lesswrong.com/lw/r1/timeless_control/)

25. [http://lesswrong.com/lw/rb/possibility\\_and\\_couldness/](http://lesswrong.com/lw/rb/possibility_and_couldness/)

26. [http://lesswrong.com/lw/r/c/the\\_ultimate\\_source/](http://lesswrong.com/lw/r/c/the_ultimate_source/)

27. Page 715, 'Joy in the Merely Real'.

28. [http://lesswrong.com/lw/r1/timeless\\_control/](http://lesswrong.com/lw/r1/timeless_control/)

29. Page 754, 'Angry Atoms'.

algorithm feels from inside<sup>30</sup>. (Being told flatly “one level emerges<sup>31</sup> from the other” just relates them by a magical<sup>32</sup> transition rule, “emergence”.)

For free will, the key is to understand how your brain computes whether you “could” do something—the algorithm that labels reachable states<sup>33</sup>. Once you understand this label, it does not appear particularly meaningless—“could” makes sense—and the label does not conflict with physics following a deterministic course. If you can *see* that, you can *see* that there is no conflict between your feeling of freedom, and deterministic physics. Indeed, I am perfectly willing to say that the feeling of freedom is correct<sup>34</sup>, when the feeling is interpreted correctly.

In the case of morality, once again there are two levels of organization, seemingly quite difficult to fit together:

On one level, there are just particles without a shred of *should*-ness built into them—just like an electron has no notion of what it “could” do—or just like a flipping coin is not uncertain of its own result<sup>35</sup>.

On another level is the ordinary morality of everyday life: moral errors, moral progress, and things you ought to do whether you want to do them or not.

And in between, the level transition question: What is this *should*-ness stuff?

Award yourself a point if you thought, “But wait, that problem isn’t quite analogous to the one of free will. With free will it was just a question of factual investigation—look at human psychology, figure out how it *does in fact* generate the feeling of freedom. But here, it won’t be enough to figure out how the mind generates its feelings of should-ness. Even after we know, we’ll be left with a remaining question—is that how we *should* calculate should-ness? So it’s not just a matter of sheer factual reductionism, it’s a moral question.”

---

<sup>30</sup>. Page 223, ‘How An Algorithm Feels From Inside’.

<sup>31</sup>. Page 158, ‘The Futility of Emergence’.

<sup>32</sup>. Page 158, ‘The Futility of Emergence’.

<sup>33</sup>. [http://lesswrong.com/lw/rb/possibility\\_and\\_couldness/](http://lesswrong.com/lw/rb/possibility_and_couldness/)

<sup>34</sup>. [http://lesswrong.com/lw/rc/the\\_ultimate\\_source/](http://lesswrong.com/lw/rc/the_ultimate_source/)

<sup>35</sup>. Page 687, ‘Probability is in the Mind’.

Award yourself *two* points if you thought, “...oh, wait, I recognize *that* pattern: It’s one of those strange loops<sup>36</sup> through the meta-level<sup>37</sup> we were talking about earlier.”

And if you’ve been reading along this whole time, you know the answer isn’t going to be, “Look at this *fundamentally* moral stuff!”

Nor even, “Sorry, morality is *mere* preference, and right-ness is just what serves you or your genes; all your moral intuitions otherwise are wrong, but I won’t explain where they come from.”

Of the art of answering<sup>38</sup> impossible<sup>39</sup> questions<sup>40</sup>, I have already said much: Indeed, vast segments of my *Overcoming Bias* posts were created with that specific hidden agenda.

The<sup>41</sup> sequence<sup>42</sup> on<sup>43</sup> anticipation<sup>44</sup> fed into Mysterious Answers to Mysterious Questions<sup>45</sup>, to prevent the Primary Catastrophic Failure of stopping on a poor answer.

The Fake Utility Functions sequence<sup>46</sup> was directed at the problem of oversimplified moral answers particularly.

The sequence on words<sup>47</sup> provided the first and basic illustration of the Mind Projection Fallacy<sup>48</sup>, the understanding of which is one of the Great Keys.

The sequence on words also showed us how to play Rationalist’s Taboo<sup>49</sup>, and Replace the Symbol with the Substance<sup>50</sup>. What is “right”, if you can’t say “good” or “desirable” or “better” or “preferable” or “moral” or “should”? What happens if you try to carry out the operation of replacing the symbol with what it stands for?

---

36. Page 887, ‘Where Recursive Justification Hits Bottom’.

37. Page 896, ‘My Kind of Reflection’.

38. Page 673, ‘Dissolving the Question’.

39. Page 679, ‘Wrong Questions’.

40. Page 682, ‘Righting a Wrong Question’.

41. Page 105, ‘Making Beliefs Pay Rent (in Anticipated Experiences)’.

42. Page 108, ‘Belief in Belief’.

43. Page 126, ‘Your Strength as a Rationalist’.

44. Page 140, ‘Guessing the Teacher’s Password’.

45. Page 154, ‘Mysterious Answers to Mysterious Questions’.

46. [http://lesswrong.com/lw/lp/fake\\_utility\\_functions/](http://lesswrong.com/lw/lp/fake_utility_functions/)

47. Page 301, ‘37 Ways That Words Can Be Wrong’.

48. Page 685, ‘Mind Projection Fallacy’.

49. Page 245, ‘Taboo Your Words’.

50. Page 248, ‘Replace the Symbol with the Substance’.

And the sequence on quantum physics<sup>51</sup>, among other purposes<sup>52</sup>, was there to teach the fine art of not *running away* from Scary and Confusing Problems<sup>53</sup>, even if others have failed to solve them, even if great minds failed to solve them for generations. Heroes screw up, time moves on, and each succeeding era gets an entirely new chance.

If you're just joining us here (Belldandy help you) then you might want to think about reading all those posts before, oh, say, tomorrow.

If you've been reading this whole time, then you should think about trying to dissolve the question<sup>54</sup> on your own, before tomorrow. It doesn't require more than 96 insights<sup>55</sup> beyond those already provided.

Next: *The Meaning of Right*.

---

51. [http://lesswrong.com/lw/r5/the\\_quantum\\_physics\\_sequence/](http://lesswrong.com/lw/r5/the_quantum_physics_sequence/)

52. [http://lesswrong.com/lw/qy/why\\_quantum/](http://lesswrong.com/lw/qy/why_quantum/)

53. [http://lesswrong.com/lw/qy/why\\_quantum/](http://lesswrong.com/lw/qy/why_quantum/)

54. Page 673, 'Dissolving the Question'.

55. [http://lesswrong.com/lw/qg/the\\_failures\\_of\\_eld\\_science/](http://lesswrong.com/lw/qg/the_failures_of_eld_science/)

## 26. The Meaning of Right<sup>1</sup>

**Continuation of:** Changing Your Metaethics<sup>2</sup>, Setting Up Metaethics<sup>3</sup>

**Followup to:** Does Your Morality Care What You Think?<sup>4</sup>, The Moral Void<sup>5</sup>, Probability is Subjectively Objective<sup>6</sup>, Could Anything Be Right?<sup>7</sup>, The Gift We Give To Tomorrow<sup>8</sup>, Rebelling Within Nature<sup>9</sup>, Where Recursive Justification Hits Bottom<sup>10</sup>, ...

(The culmination of a *long* series of *Overcoming Bias* posts; if you start here, I accept no responsibility for any resulting confusion, misunderstanding, or unnecessary angst<sup>11</sup>.)

What *is* morality? What does the word “should”, *mean*? The many pieces<sup>12</sup> are in place: This question I shall now dissolve<sup>13</sup>.

The key—as it has always been, in my experience so far—is to understand how a certain cognitive algorithm feels from inside<sup>14</sup>. Standard procedure for righting a wrong question<sup>15</sup>: If you don’t know what right-ness is, then take a step beneath and ask how your brain labels things “right”.

It is not the *same* question—it has no moral aspects to it, being strictly a matter of fact and cognitive science. But it is an *illuminating* question. Once we know how our brain labels things “right”, perhaps we shall find it easier, afterward, to ask what is really and truly *right*.

But with that said—the easiest way to begin investigating *that* question, will be to jump back up to the level of morality and ask what *seems* right. And if that seems like too much recursion, get

---

1. [http://lesswrong.com/lw/sm/the\\_meaning\\_of\\_right/](http://lesswrong.com/lw/sm/the_meaning_of_right/)

2. Page 962, ‘Changing Your Metaethics’.

3. Page 967, ‘Setting Up Metaethics’.

4. Page 955, ‘Does Your Morality Care What You Think?’.

5. Page 856, ‘The Moral Void’.

6. Page 915, ‘Probability is Subjectively Objective’.

7. Page 935, ‘Could Anything Be Right?’.

8. Page 927, ‘The Gift We Give To Tomorrow’.

9. Page 907, ‘Rebelling Within Nature’.

10. Page 887, ‘Where Recursive Justification Hits Bottom’.

11. Page 940, ‘Existential Angst Factory’.

12. Page 967, ‘Setting Up Metaethics’.

13. Page 673, ‘Dissolving the Question’.

14. Page 223, ‘How An Algorithm Feels From Inside’.

15. Page 682, ‘Righting a Wrong Question’.

used to it—the other 90% of the work lies in handling recursion properly.

(Should you find your grasp on meaningfulness wavering, at any time following, check *Changing Your Metaethics*<sup>16</sup> for the appropriate prophylactic.)

So! In order to investigate how the brain labels things “right”, we are going to *start out* by talking about what is right. That is, we’ll start out wearing our *morality-goggles*, in which we consider morality-as-morality and talk about moral questions directly. As opposed to wearing our *reduction-goggles*, in which we talk about cognitive algorithms and mere physics. Rigorously distinguishing between these two views is the first step toward mating them together.

As a first step, I offer this observation, on the level of morality-as-morality: Rightness is contagious backward in time.

Suppose there is a switch, currently set to OFF, and it is *morally desirable* for this switch to be flipped to ON. Perhaps the switch controls the emergency halt on a train bearing down on a child strapped to the railroad tracks, this being my canonical example. If this is the case, then, *ceteris paribus* and presuming the absence of exceptional conditions or further consequences that were not explicitly specified, we may consider it *right* that this switch should be flipped.

If it is right to flip the switch, then it is right to pull a string that flips the switch. If it is good to pull a string that flips the switch, it is right and proper to press a button that pulls the string: Pushing the button seems to have more *should-ness* than not pushing it.

It seems that—all else being equal, and assuming no other consequences or exceptional conditions which were not specified—value flows backward along arrows of causality.

Even in deontological moralities, if you’re *obligated* to save the child on the tracks, then you’re *obligated* to press the button. Only very primitive AI systems have motor outputs controlled by strictly local rules that don’t model the future *at all*. Duty-based or virtue-based ethics are only *slightly* less consequentialist than consequentialism. It’s hard to say whether moving your arm left or right is more virtuous without talking about what happens next.

---

<sup>16</sup> Page 962, ‘*Changing Your Metaethics*’.

Among my readers, there may be some who presently assert—though I hope to persuade them otherwise—that the life of a child is of no value to them. If so, they may substitute anything else that they prefer, at the end of the switch, and ask if they should press the button.

But I also suspect that, among my readers, there are some who wonder if the *true* morality might be something quite different<sup>17</sup> from what is presently believed among the human kind. They may find it imaginable—plausible?—that human life is of no value, or negative value. They may wonder if the goodness of human happiness, is as much a self-serving delusion as the justice of slavery.

I myself<sup>18</sup> was once numbered among these skeptics, because I was always very suspicious<sup>19</sup> of anything that looked self-serving.

Now here's a little question I never thought to ask, during those years when I thought I knew nothing about morality<sup>20</sup>:

Could make sense to have a morality in which, if we *should* save the child from the train tracks, then we *should not* flip the switch, *should* pull the string, and *should not* push the button, so that, finally, we do not push the button?

Or perhaps someone says that it is better to save the child, than to not save them; but doesn't see why anyone would think this implies it is better to press the button than not press it. (Note the resemblance to the Tortoise who denies *modus ponens*<sup>21</sup>.)

17. Page 935, ‘Could Anything Be Right?’.

18. Page 167, ‘My Wild and Reckless Youth’.

19. Page 907, ‘Rebelling Within Nature’.

20. Page 935, ‘Could Anything Be Right?’.

21. Page 861, ‘Created Already In Motion’.

It seems imaginable, to at least some people, that entirely different things could be *should*. It didn't seem nearly so imaginable, at least to me, that *should*-ness could fail to flow backward in time. When I was trying to question everything else, that thought simply did not occur to me.

Can you question it? Should you?

Every now and then, in the course of human existence, we question what *should* be done and what is *right* to do, what is *better* or *worse*; others come to us with assertions along these lines, and we question them, asking "Why is it right?" Even when we believe a thing is right (because someone told us that it is, or because we wordlessly feel that it is) we may still question why it is right.

*Should*-ness, it seems, flows backward in time. This gives us one way to question why or whether a particular event has the *should*-ness property. We can look for some *consequence* that has the *should*-ness property. If so, the *should*-ness of the original event seems to have been plausibly proven or explained.

Ah, but what about the consequence—why is *it* should? Someone comes to you and says, "You should give me your wallet, because then I'll have your money, and I should have your money." If, at this point, you stop asking questions<sup>22</sup> about *should*-ness, you're vulnerable to a moral mugging.

So we keep asking the next question. Why should we press the button? To pull the string. Why should we pull the string? To flip the switch. Why should we flip the switch? To pull the child from the railroad tracks. Why pull the child from the railroad tracks? So that they live. Why should the child live?

Now there are people who, caught up in the enthusiasm, go ahead and answer that question in the same style: for example, "Because the child might eventually grow up and become a trade partner with you," or "Because you will gain honor in the eyes of others," or "Because the child may become a great scientist and help achieve the Singularity," or some such. But even if we were to answer in this style, it would only beg the next question.

---

<sup>22</sup> Page 175, 'Explain/Worship/Ignore?'

Even if you try to have a chain of *should* stretching into the infinite future—a trick I've yet to see anyone try to pull, by the way, though I may be only ignorant of the breadths of human folly—then you would simply ask “Why that chain<sup>23</sup> rather than some other?”

Another way that something can be *should*, is if there's a general rule that makes it *should*. If your belief pool starts out with the general rule “All children X: It is better for X to live than to die”, then it is quite a short step to “It is better for Stephanie to live than to die”. Ah, but why save all children? Because they may all become trade partners or scientists? But then where did *that* general rule come from?

If *should*-ness only comes from *should*-ness—from a *should*-consequence, or from a *should*-universal—then how does anything end up *should* in the first place?

Now human beings have argued these issues for thousands of years and maybe much longer. We do not hesitate to continue arguing when we reach a terminal value<sup>24</sup> (something that has a charge of *should*-ness independently of its consequences). We just go on arguing about the universals.

I usually take, as my archetypal example, the undoing of slavery: Somehow, slaves' lives went from having no value to having value. Nor do I think that, back at the dawn of time, anyone was even trying to argue that slaves were better off being slaves (as it would be latter argued). They'd probably have looked at you like you were crazy if you even tried. Somehow, we got from there, to here...

And some of us would even hold this up as a case of moral progress<sup>25</sup>, and look at our ancestors as having made a *moral error*. Which seems easy enough to describe in terms of *should*-ness: Our ancestors *thought* that they should enslave defeated enemies, but they were mistaken.

But all our philosophical arguments ultimately seem to ground in statements that no one has bothered to justify—except perhaps to plead that they are *self-evident*, or that any *reasonable* mind must surely agree, or that they are *a priori* truths, or some such. Perhaps, then, *all* our moral beliefs are as erroneous as that old bit about slav-

<sup>23</sup> [http://lesswrong.com/lw/rc/the\\_ultimate\\_source/](http://lesswrong.com/lw/rc/the_ultimate_source/)

<sup>24</sup> [http://lesswrong.com/lw/l4/terminal\\_values\\_and\\_instrumental\\_values/](http://lesswrong.com/lw/l4/terminal_values_and_instrumental_values/)

<sup>25</sup> Page 924, ‘Whither Moral Progress?’.

ery? Perhaps we have entirely misperceived the flowing streams of *should*?

So I once believed was plausible; and one of the arguments I wish I could go back and say to myself, is, “If you know nothing at all about should-ness<sup>26</sup>, then how do you know that the procedure, ‘Do whatever Emperor Ming says’ is not the entirety of should-ness? Or even worse, perhaps, the procedure, ‘Do whatever maximizes inclusive genetic fitness’ or ‘Do whatever makes you personally happy.’” The point here would have been to make my past self see that in *rejecting* these rules, he was asserting a kind of knowledge—that to say, “This is *not* morality,” he must reveal that, despite himself, he knows something about morality or meta-morality. Otherwise, the procedure “Do whatever Emperor Ming says” would seem *just* as plausible, as a guiding principle, as his current path of “Rejecting things that seem unjustified.” Unjustified—according to what criterion of *justification*? Why trust the principle that says that moral statements need to be justified, if you know nothing at all about morality?

What indeed would distinguish, *at all*, the question “What is right?” from “What is wrong?”

What is “right”, if you can’t say<sup>27</sup> “good” or “desirable” or “better” or “preferable” or “moral” or “should”? What happens if you try to carry out the operation of replacing the symbol with what it stands for<sup>28</sup>?

If you’re guessing that I’m trying to inveigle you into letting me say: “Well, there are just some things that are baked into the *question*, when you start asking questions about *morality*, rather than wakalixes or toaster ovens”, then you would be right. I’ll be making use of that later, and, yes, will address “But why *should* we ask that question?”

*Okay, now: morality-goggles off, reduction-goggles on.*

Those who remember Possibility and Could-ness<sup>29</sup>, or those familiar with simple search techniques in AI, will realize that the “should” label is behaving like the inverse of the “could” label, which we previously analyzed in terms of “reachability”. Reach-

---

26. Page 935, ‘Could Anything Be Right?’.

27. Page 245, ‘Taboo Your Words’.

28. Page 248, ‘Replace the Symbol with the Substance’.

29. [http://lesswrong.com/lw/rb/possibility\\_and\\_couldness/](http://lesswrong.com/lw/rb/possibility_and_couldness/)

bility spreads *forward* in time: if I could reach the state with the button pressed, I could reach the state with the string pulled; if I could reach the state with the string pulled, I could reach the state with the switch flipped.

Where the “could” label and the “should” label collide, the algorithm produces a plan.

Now, as I say this, I suspect that at least some readers may find themselves fearing that I am about to reduce *should*-ness to a *mere* artifact of a way that a planning system feels from inside. Once again I urge you to check *Changing Your Metaethics*<sup>30</sup>, if this starts to happen. Remember above all the Moral Void<sup>31</sup>: Even if there were no morality, you could still choose to help people rather than hurt them. This, above all, holds in place what you hold precious, while your beliefs about the nature of morality change.

I do not intend, with this post, to take away anything of value; it will all be given back before the end.

Now this algorithm is not very sophisticated, as AI algorithms go, but to apply it in full generality—to learned information, not just ancestrally encountered, genetically programmed situations—is a rare thing among animals. Put a food reward in a transparent box. Put the matching key, which looks unique and uniquely corresponds to that box, in another transparent box. Put the unique key to *that* box in another box. Do this with five boxes. Mix in another sequence of five boxes that doesn’t lead to a food reward. Then offer a choice of two keys, one of which starts the sequence of five boxes leading to food, one of which starts the sequence leading nowhere.

Chimpanzees can learn to do this, but so far as I know, no non-primate species can pull that trick.

And as smart as chimpanzees are, they are not quite as good as humans at inventing plans—plans such as, for example, planting in the spring to harvest in the fall.

---

<sup>30</sup>. Page 962, ‘*Changing Your Metaethics*’.

<sup>31</sup>. Page 856, ‘*The Moral Void*’.

So what else are humans doing, in the way of planning?

It is a general observation that natural selection seems to *reuse* existing complexity, rather than creating things from scratch, whenever it *possibly* can—though not always in the *same way* that a human engineer would. It is a function of the enormous time<sup>32</sup> required for evolution to create machines with many interdependent parts, and the vastly shorter time required to create a mutated copy of something already evolved.

What else are humans doing? Quite a bit, and some of it I don't understand—there are plans humans make, that no modern-day AI can.

But *one* of the things we are doing, is reasoning about “right-ness” the same way we would reason about any other observable property.

Are animals with bright colors often poisonous? Does the delicious nid-nut grow only in the spring? Is it usually a good idea to take with a waterskin on long hunts?

It seems that Martha and Fred have an obligation to take care of their child, and Jane and Bob are obligated to take care of their child, and Susan and Wilson have a duty to care for their child. Could it be that parents in general must take care of their children?

By representing right-ness as an attribute of objects, you can recruit a whole previously evolved system that reasons about the attributes of objects. You can save quite a lot of planning time, if you decide (based on experience) that *in general* it is a good idea to take a waterskin on hunts, from which it follows that it must be a good idea to take a waterskin on hunt #342.

Is this damnable for a Mind Projection Fallacy<sup>33</sup>—treating properties of the mind as if they were out there in the world?

Depends on how you look at it.

This business of, “It's been a good idea to take waterskins on the last three hunts, maybe it's a good idea in general, if so it's a good idea to take a waterskin on this hunt”, does seem to *work*.

Let's say that your mind, faced with any countable set of objects, automatically and perceptually tagged them with their remainder modulo 5. If you saw a group of 17 objects, for example, they would

---

32. [http://lesswrong.com/lw/kt/evolutions\\_are\\_stupid\\_but\\_work\\_anyway/](http://lesswrong.com/lw/kt/evolutions_are_stupid_but_work_anyway/)

33. Page 685, ‘Mind Projection Fallacy’.

look *remainder-2-ish*. Though, if you didn't have any notion of *what* your neurons were doing, and perhaps no notion of modulo arithmetic, you would only see that the group of 17 objects had the same *remainder-ness* as a group of 2 objects. You might not even know how to count—your brain doing the whole thing automatically, subconsciously and neurally—in which case you would just have five different words for the *remainder-ness* attributes that we would call 0, 1, 2, 3, and 4.

If you look out upon the world you see, and guess that *remainder-ness* is a separate and additional attribute of things—like the attribute of having an electric charge—or like a tiny little XML tag hanging off of things—then you will be wrong. But this does not mean it is nonsense to talk about *remainder-ness*, or that you must automatically commit the Mind Projection Fallacy<sup>34</sup> in doing so. So long as you've got a well-defined way to compute a property, it can have a well-defined output and hence an empirical truth condition.

If you're looking at 17 objects, then their *remainder-ness* is, indeed and truly, 2, and not 0, 3, 4, or 1. If I tell you, "Those red things you told me to look at are *remainder-2-ish*", you have indeed been told a falsifiable and empirical property of those red things. It is just not a separate, additional, physically existent attribute.

And as for reasoning *about* derived properties, and which other inherent or derived properties they correlate to—I don't see anything inherently fallacious about that.

One may notice, for example, that things which are 7 modulo 10 are often also 2 modulo 5. *Empirical* observations of this sort play a large role in mathematics, suggesting theorems to prove. (See Polya's *How To Solve It*.)

Indeed, virtually all the experience we have, is derived by complicated neural computations from the raw physical events impinging on our sense organs. By the time you *see* anything, it has been extensively processed by the retina, lateral geniculate nucleus, visual cortex, parietal cortex, and temporal cortex, into a very complex sort of derived computational property.

If you thought of a property like *redness* as residing strictly *in an apple*, you would be committing the Mind Projection Fallacy. The

---

<sup>34</sup> Page 685, 'Mind Projection Fallacy'.

apple's surface has a reflectance which sends out a mixture of wavelengths that impinge on your retina and are processed with respect to ambient light to extract a summary color of *red*... But if you tell me that the apple is red, rather than green, and make no claims as to whether this is an ontologically fundamental physical attribute of the apple, then I am quite happy to agree with you.

So as long as there is a stable computation involved, or a stable process—even if you can't consciously verbalize the specification—it often makes a great deal of sense to talk about properties that are not fundamental. And reason about them, and remember where they have been found in the past, and guess where they will be found next.

(In retrospect, that should have been a separate post in the Reductionism sequence. “Derived Properties”, or “Computational Properties” maybe. Oh, well; I promised you morality this day, and this day morality you shall have.)

Now let's say we want to make a little machine, one that will save the lives of children. (This enables us to save more children than we could do without a machine, just like you can move more dirt with a shovel than by hand.) The machine will be a planning machine, and it will reason about events that may or may not have the property, *leads-to-child-living*.

A simple planning machine would just have a pre-made model of the environmental process. It would search forward from its actions, applying a label that we might *call* “reachable-from-actionness”, but which might as well say “Xybliz” internally for all that it matters to the program. And it would search backward from scenarios, situations, in which the child lived, labeling these “leads-to-child-living”. If situation X leads to situation Y, and Y has the label “leads-to-child-living”—which might just be a little flag bit, for all the difference it would make—then X will inherit the flag from Y. When the two labels meet in the middle, the leads-to-child-living flag will quickly trace down the stored path of reachability, until finally some particular sequence of actions ends up labeled “leads-to-child-living”. Then the machine automatically executes those actions—that's just what the machine does.

Now this machine is not complicated enough to feel existential angst. It is not complicated enough to commit the Mind Projection Fallacy. It is not, in fact, complicated enough to *reason abstractly* about the property “leads-to-child-living-ness”. The machine—as specified so far—does not notice if the action “jump in the air” turns out to always have this property, or never have this property. If “jump in the air” always led to situations in which the child lived, this could greatly simplify future planning—but only if the machine were sophisticated enough to notice this fact and use it.

If it is a fact that “jump in the air” “leads-to-child-living-ness”, this fact is composed of empirical truth and logical truth. It is an *empirical* truth that if the world is such that if you perform the (ideal abstract) algorithm “trace back from situations where the child lives”, then it will be a *logical* truth about the output of this (ideal abstract) algorithm that it labels the “jump in the air” action.

(You cannot always define this fact in *entirely* empirical terms, by looking for the physical real-world coincidence of jumping and child survival. It might be that “stomp left” *also* always saves the child, and the machine in fact stomps left. In which case the fact that jumping in the air *would have* saved the child, is a counterfactual extrapolation<sup>35</sup>.)

*Okay, now we’re ready to bridge the levels.*

As you must surely have guessed by now, this *should-ness* stuff is how the human decision algorithm feels from inside<sup>36</sup>. It is not an extra, physical, ontologically fundamental attribute hanging off of events like a tiny little XML tag.

But it is a *moral* question what we should do about that—how we should react to it.

To adopt an attitude of complete nihilism, because we *wanted* those tiny little XML tags, and they’re *not physically there*, strikes me as the wrong move. It is like supposing that the absence of an XML tag, equates to the XML tag *being there*, saying in its tiny brackets *what value we should attach*, and having value zero. And then this value zero, in turn, equating to a moral imperative to wear black, feel awful, write gloomy poetry, betray friends, and commit suicide.

---

35. Page 944, ‘Can Counterfactuals Be True?’.

36. Page 223, ‘How An Algorithm Feels From Inside’.

No.

So what would I say instead?

The force behind my answer is contained in The Moral Void<sup>37</sup> and The Gift We Give To Tomorrow<sup>38</sup>. I would try to save lives “even if there were no morality”, as it were.

And it seems like an awful shame to—after so many millions and hundreds of millions of years of evolution—after the moral miracle<sup>39</sup> of so much cutthroat genetic competition producing intelligent minds that love, and hope, and appreciate beauty, and create beauty—after coming so far, to throw away the Gift of morality, *just because our brain happened to represent morality in such fashion as to potentially mislead us when we reflect on the nature of morality*.

This little accident of the Gift doesn’t seem like a good reason to throw away the Gift; it certainly isn’t a inescapable logical justification for wearing black.

Why not keep the Gift, but adjust the way we reflect on it?

So here’s my metaethics:

I earlier asked,

What is “right”, if you can’t say<sup>40</sup> “good” or “desirable” or “better” or “preferable” or “moral” or “should”? What happens if you try to carry out the operation of replacing the symbol with what it stands for<sup>41</sup>?

I answer that if you try to replace the symbol “should” with *what it stands for*, you end up with quite a large sentence.

For the much simpler save-life machine, the “should” label stands for leads-to-child-living-ness.

For a human this is a much huger blob of a computation that looks like, “Did everyone survive? How many people are happy? Are people in control of their own lives? ...” Humans have complex emotions, have many values—the thousand shards of desire, the godshatter of natural selection<sup>42</sup>. I would say, by the way, that the

---

37. Page 856, ‘The Moral Void’.

38. Page 927, ‘The Gift We Give To Tomorrow’.

39. Page 927, ‘The Gift We Give To Tomorrow’.

40. Page 245, ‘Taboo Your Words’.

41. Page 248, ‘Replace the Symbol with the Substance’.

42. [http://lesswrong.com/lw/l3/thou\\_art\\_godshatter/](http://lesswrong.com/lw/l3/thou_art_godshatter/)

huge blob of a computation is not just my present terminal values (which I don't really *have*—I am not a consistent expected utility maximizers); the huge blob of a computation includes the specification of those moral arguments, those justifications, that would sway me if I heard them. So that I can regard my present values, as an approximation to the ideal morality that I would have if I heard all the arguments, to whatever extent such an extrapolation is coherent<sup>43</sup>.

No one can write down their big computation; it is not just too large, it is also unknown to its user. No more could you print out a listing of the neurons in your brain. You never *mention* your big computation—you only *use* it, every hour of every day.

Now why might one *identify* this enormous abstract computation, with what-is-right?

If you identify rightness with this *huge computational property*, then moral judgments are subjunctively objective<sup>44</sup> (like math), subjectively objective<sup>45</sup> (like probability), and capable of being true<sup>46</sup> (like counterfactuals).

You will find yourself saying, “If I wanted to kill someone—even if I thought it was right to kill someone—that wouldn’t make it right.” Why? Because what is *right* is a huge computational property—an *abstract computation—not tied to the state of anyone’s brain, including your own brain.*

This distinction was introduced earlier in 2-Place and 1-Place Words<sup>47</sup>. We can treat the word “sexy” as a 2-place function that goes out and hoovers up someone’s sense of sexiness, and then eats an object of admiration. Or we can treat the word “sexy” as *meaning* a 1-place function, a *particular* sense of sexiness, like Sexiness\_20934, that only accepts one argument, an object of admiration.

Here we are treating morality as a *1-place function*. It does not accept a person as an argument, spit out whatever cognitive algorithm they use to choose between actions, and then apply that

43. <http://www.google.com/>

*url?sa=t&ct=res&cd=9&url=http%3A%2F%2Fwww.singinst.org%2Fupload%2FCEV.html&ei=RlmOSPCCZHosA2gQ4J5Ruf018U5442ox7A&sig2=EHQy6Qoc9cdq9qTp2kRqow*

44. Page 948, ‘Math is Subjunctively Objective’.

45. Page 915, ‘Probability is Subjectively Objective’.

46. Page 944, ‘Can Counterfactuals Be True?’.

47. Page 849, ‘2-Place and 1-Place Words’.

algorithm to the situation at hand. When I say *right*, I mean a certain *particular* 1-place function that just asks, “Did the child live? Did anyone else get killed? Are people happy? Are they in control of their own lives? Has justice been served?” ... and so on through many, many other elements of rightness. (And perhaps those arguments that might persuade me otherwise, which I have not heard.)

Hence the notion, “Replace the symbol with what it stands for.”

Since what’s *right* is a 1-place function, if I subjunctively imagine a world in which someone has slipped me a pill that makes me want to kill people, then, in this subjunctive world, it is not *right* to kill people. That’s not merely because I’m judging with my current brain. It’s because when I say *right*, I am referring to a 1-place function. Rightness doesn’t go out and Hoover up the current state of my brain, in this subjunctive world, before producing the judgment “Oh, wait, it’s now okay to kill people.” When I say *right*, I don’t *mean* “that which my future self wants”, I *mean* the function that looks at a situation and asks, “Did anyone get killed? Are people happy? Are they in control of their own lives? ...”

And once you’ve defined a particular abstract computation that says what is *right*—or even if you haven’t defined it, and it’s computed in some part of your brain you can’t perfectly print out, but the computation is *stable*—more or less—then as with any other derived property, it makes sense to speak of a moral judgment being *true*. If I say that today was a good day, you’ve learned something empirical and falsifiable about my day—if it turns out that actually my grandmother died, you will suspect that I was originally lying.

The apparent objectivity of morality has just been explained—and *not* explained *away*. For indeed, if someone slipped me a pill that made me want to kill people, nonetheless, it would not be *right* to kill people. Perhaps I would actually kill people, in that situation—but that is because something other than morality would be controlling my actions.

Morality is not just subjunctively objective, but subjectively objective. I experience it as something I cannot change. Even after I know that it’s myself who computes this 1-place function, and not a rock somewhere—even after I know that I will not find any star or mountain that computes this function, that only upon me is it written—even so, I find that I wish to save lives, and that even if I

could change this by an act of will, I would not choose to do so. I do not wish to reject joy, or beauty, or freedom. What else would I do instead? I do not wish to reject the Gift that natural selection accidentally barfed into me. This is the principle of The Moral Void<sup>48</sup> and The Gift We Give To Tomorrow<sup>49</sup>.

Our origins may seem unattractive, our brains untrustworthy.

But love has to enter the universe somehow, starting from non-love, or love cannot enter time<sup>50</sup>.

And if our brains are untrustworthy, it is only our own brains that say so. Do you sometimes think that human beings are not very nice? Then it is you, a human being, who says so. It is you, a human being, who judges that human beings could do better<sup>51</sup>. You will not find such written upon the stars or the mountains: they are not minds, they cannot think.

In this, of course, we find a justificational strange loop through the meta-level<sup>52</sup>. Which is unavoidable so far as I can see—you can't argue morality, or any kind of goal optimization, into a rock. But note the exact structure of this strange loop: *there is no general moral principle which says that you should do what evolution programmed you to do*. There is, indeed, no general principle to trust your moral intuitions! You can find a moral intuition within yourself, describe it—quote it—consider it deliberately and in the full light of your entire morality, and reject it, on grounds of other arguments. What counts as an argument is also built into the rightness-function.

Just as, in the strange loop of rationality, there is no general principle in rationality to trust your brain, or to believe what evolution programmed you to believe—but indeed, when you ask which parts of your brain you need to rebel<sup>53</sup> against, you do so using your current brain. When you ask whether the universe is simple, you can consider the *simple* hypothesis that the universe's apparent simplicity is explained by its actual simplicity.

Rather than trying to unwind ourselves into rocks, I proposed that we should use the *full strength* of our current rationality, in re-

48. Page 856, ‘The Moral Void’.

49. Page 927, ‘The Gift We Give To Tomorrow’.

50. Page 927, ‘The Gift We Give To Tomorrow’.

51. [http://lesswrong.com/lw/b8/tsuyoku\\_naritai\\_i\\_want\\_to\\_become\\_stronger/](http://lesswrong.com/lw/b8/tsuyoku_naritai_i_want_to_become_stronger/)

52. Page 887, ‘Where Recursive Justification Hits Bottom’.

53. Page 907, ‘Rebelling Within Nature’.

flecting upon ourselves—that no part of ourselves be immune from examination, and that we use all of ourselves that we currently believe in to examine it.

You would do the same thing with morality; if you consider that a part of yourself might be considered harmful, then use your *best* current guess at what is *right*, your full moral strength, to do the considering. Why *should* we want to unwind ourselves to a rock? Why *should* we do less than our best, when reflecting? You can't unwind past Occam's Razor, modus ponens, or morality *and it's not clear why you should try*.

For any part of rightness, you can always imagine another part that overrides it—it would not be right to drag the child from the train tracks, if this resulted in everyone on Earth becoming unable to love—or so I would judge. For every part of rightness you examine, you will find that it cannot be the sole and perfect and only criterion of rightness. This may lead to the incorrect inference that there is something beyond, some perfect and only criterion from which all the others are derived—but that does not follow. The whole is the sum of the parts. We ran into an analogous situation with free will, where no part of ourselves seems perfectly decisive.<sup>54</sup>

The classic dilemma for those who would trust their moral intuitions, I believe, is the one who says: “Interracial marriage is repugnant—it disgusts me—and that is my moral intuition!” I reply, “There is no general rule to obey your intuitions. You just *mentioned* intuitions, rather than *using* them. Very few people have legitimate cause to *mention* intuitions—Friendly AI programmers, for example, delving into the cognitive science of things, have a legitimate reason to mention them. Everyone else just has ordinary moral arguments, in which they *use* their intuitions, for example, by saying, ‘An interracial marriage doesn’t hurt anyone, if both parties consent’. I do not say, ‘And I have an intuition that anything consenting adults do is right, and all intuitions must be obeyed, therefore I win.’ I just offer up that argument, and any others I can think of, to weigh in the balance.”

Indeed, evolution that made us cannot be trusted<sup>55</sup>—so there is no general principle to trust it! Rightness is not defined in terms of automatic correspondence to any possible decision we actual-

---

54. [http://lesswrong.com/lw/rc/the\\_ultimate\\_source/](http://lesswrong.com/lw/rc/the_ultimate_source/)

55. Page 907, ‘Rebelling Within Nature’.

ly make—so there's no general principle that says you're infallible! Just do what is, ahem, *right*—to the best of your ability to weigh the arguments you have heard, and ponder the arguments you may not have heard.

If you were hoping to have a perfectly trustworthy system, or to have been created in correspondence with a perfectly trustworthy morality—well, I can't give *that* back to you; but even most religions don't try that one. Even most religions have the human psychology containing elements of sin, and even most religions don't *actually* give you an effectively executable and perfect procedure, though they may tell you “Consult the Bible! It always works!”

If you hoped to find a source of morality outside humanity—well, I can't give that back, but I can ask once again: Why would you even *want* that?<sup>56</sup> And what good would it do? Even if there were some great light in the sky—something that could tell us, “Sorry, happiness is bad for you, pain is better, now get out there and kill some babies!”—it would still be your own decision to follow it. You cannot evade responsibility.

There isn't enough mystery *left* to justify *reasonable doubt* as to whether the causal origin of morality is something outside humanity. We have evolutionary psychology. We know where morality came from. We pretty much know how it works, in broad outline at least. We know there are no little XML value tags on electrons (and indeed, even if you found them, why *should* you pay attention to what is written there?)

If you hoped that morality would be universalizable—sorry, that one I *really* can't give back. Well, unless we're just talking about humans. Between neurologically intact *humans*, there is indeed much cause to hope for overlap and coherence; and a great and reasonable doubt as to whether any present disagreement is *really* irresolvable, even it seems to be about “values”. The obvious reason for hope is the psychological unity of humankind<sup>57</sup>, and the intuitions of symmetry, universalizability, and simplicity that we execute in the course of our moral arguments. (In retrospect, I should have done a post on Interpersonal Morality before this...)

---

<sup>56</sup>. Page 856, ‘The Moral Void’.

<sup>57</sup>. [http://lesswrong.com/lw/rl/the\\_psychological\\_unity\\_of\\_humankind/](http://lesswrong.com/lw/rl/the_psychological_unity_of_humankind/)

If I tell you that three people have found a pie and are arguing about how to divide it up<sup>58</sup>, the thought “Give one-third of the pie to each” is bound to occur to you—and if the three people are humans, it’s bound to occur to them, too. If one of them is a psychopath and insists on getting the whole pie, though, there may be nothing for it but to say: “Sorry, fairness is not ‘what everyone thinks is fair’, fairness is everyone getting a third of the pie<sup>59</sup>”. You might be able to resolve the remaining disagreement by politics and game theory, short of violence—but that is not the same as coming to agreement on values. (Maybe you could persuade the psychopath that taking a pill to be more human, if one were available, would make them happier? Would you be justified in forcing them to swallow the pill? These get us into stranger waters that deserve a separate post.)

If I define rightness to include the space of arguments that move me, then when you and I argue about *what is right*, we are arguing our *approximations* to what we would come to believe if we knew all empirical facts and had a million years to think about it—and that might be a lot closer than the present and heated argument. Or it might not. This gets into the notion of ‘construing an extrapolated volition’ which would be, again, a separate post.

But if you were stepping outside the human and hoping for moral arguments that would persuade any possible mind, even a mind that just wanted to maximize the number of paperclips in the universe, then sorry—the space of possible mind designs is too large<sup>60</sup> to permit universally compelling arguments<sup>61</sup>. You are better off treating your intuition that your moral arguments ought to persuade others, as applying only to other humans who are more or less neurologically intact. Trying it on human psychopaths would be dangerous, yet perhaps possible. But a paperclip maximizer is just not the sort of mind that would be moved by a *moral* argument. (This will definitely be a separate post.)

Once, in my wild and reckless youth<sup>62</sup>, I tried dutifully—I thought it was my duty—to be ready and willing to follow the dic-

58. Page 864, ‘The Bedrock of Fairness’.

59. Page 864, ‘The Bedrock of Fairness’.

60. [http://lesswrong.com/lw/rm/the\\_design\\_space\\_of\\_mindsingeneral/](http://lesswrong.com/lw/rm/the_design_space_of_mindsingeneral/)

61. Page 844, ‘No Universally Compelling Arguments’.

62. Page 167, ‘My Wild and Reckless Youth’.

tates of a great light in the sky, an external objective morality, when I discovered it. I questioned everything, even altruism toward human lives, even the value of happiness. Finally I realized that there was no foundation but humanity—no evidence pointing to even a reasonable doubt that there was anything else—and indeed I shouldn't even *want* to hope for anything else—and indeed would have no moral cause to follow the dictates of a light in the sky, even if I found one.

I didn't get back *immediately* all the pieces of myself that I had tried to deprecate—it took time for the realization “There is nothing else” to sink in. The notion that humanity could just... you know... live and have fun... seemed much too good to be true, so I mistrusted it. But eventually, it sank in that there really *was* nothing else to take the place of beauty. And then I got it back.

So you see, it all really *does* add up to moral normality, very exactly in fact. You go on with the same morals as before, and the same moral arguments as before. There is no sudden Grand Overlord Procedure to which you can appeal to get a perfectly trustworthy answer. You don't know, cannot print out, the great rightness-function; and even if you could, you would not have enough computational power to search the entire specified space of arguments that might move you. You will just have to argue it out.

I suspect that a fair number of those who propound metaethics do so in order to have it add up to some new and unusual moral—else why would they bother? In my case, I bother because I am a Friendly AI programmer and I have to make a physical system outside myself do what's right; for which purpose metaethics becomes very important indeed. But for the most part, the effect of my proffered metaethic is threefold:

- Anyone worried that reductionism drains the meaning from existence can stop worrying;
- Anyone who was rejecting parts of their human existence based on strange metaethics—i.e., “Why should I care about others, if that doesn't help me maximize my inclusive genetic fitness?”—can welcome back all the parts of themselves that they once exiled.
- You can stop arguing about metaethics, and go back to whatever ordinary moral argument you were having before then. This knowledge will help you avoid metaethical

*mistakes* that mess up moral arguments, but you can't actually use it to *settle debates* unless you can build a Friendly AI.

And, oh yes—*why* is it *right* to save a child's life?

Well... you could ask “Is this event that just happened, right?” and find that the child had survived, in which case you would have discovered the nonobvious empirical fact about the world, that it had come out right.

Or you could start out already knowing a complicated state of the world, but still have to apply the rightness-function to it in a nontrivial way—one involving a complicated moral argument, or extrapolating consequences into the future—in which case you would learn the nonobvious logical / computational fact that rightness, applied to this situation, yielded thumbs-up.

In both these cases, there are nonobvious facts to learn, which seem to *explain* why what just happened is *right*.

But if you ask “Why is it good to be happy?” and then replace the symbol ‘good’ with what it stands for, you’ll end up with a question like “Why does happiness match {happiness + survival + justice + individuality + ...}?” This gets computed so fast, that it scarcely seems like there’s anything there to be explained. It’s like asking “Why does  $4 = 4$ ?” instead of “Why does  $2 + 2 = 4$ ?”

Now, I bet that feels quite a bit like what happens when I ask you: “Why is happiness good?”

Right?

And that’s also my answer to Moore’s Open Question. Why is this big function I’m talking about, *right*? Because when I say “that big function”, and you say “right”, we are dereferencing two different pointers to the same unverbalizable abstract computation. I mean, that big function I’m talking about, happens to be the same thing that labels things *right* in your own brain. You might reflect on the pieces of the quotation of the big function, but you would start out by using your sense of *right-ness* to do it. If you had the perfect empirical knowledge to taboo both “that big function” and “right”, substitute what the pointers stood for, and write out the full enormity of the resulting sentence, it would come out as... sorry, I can’t resist this one... A=A.

## 27. Interpersonal Morality<sup>1</sup>

### Followup to: The Bedrock of Fairness<sup>2</sup>

Every time I wonder if I really need to do so much prep work to explain an idea, I manage to forget some minor thing and a dozen people promptly post objections.

In this case, I seem to have forgotten to cover the topic of how morality applies to more than one person at a time.

Stop laughing, it's not quite as dumb an oversight as it sounds. Sort of like how some people argue that macroeconomics should be constructed from microeconomics, I tend to see interpersonal morality as constructed from personal morality. (And definitely not the other way around!)

In "The Bedrock of Fairness"<sup>3</sup> I offered a situation where three people discover a pie, and one of them *insists* that they want half. This is actually toned down from an older dialogue where five people discover a pie, and one of them—regardless of any argument offered—insists that they want the *whole* pie.

Let's consider the latter situation: Dennis wants the whole pie. Not only that, Dennis says that it is "fair" for him to get the whole pie, and that the "right" way to *resolve this group disagreement* is for him to get the whole pie; and he goes on saying this no matter what arguments are offered him.

This group is not going to agree, no matter what. But I would, nonetheless, say that the *right* thing to do, the *fair* thing to do, is to give Dennis one-fifth of the pie—the other four combining to hold him off by force, if necessary, if he tries to take more.

A terminological note:

In this series of posts I have been using "morality" to mean something more like "the sum of all values and valuation rules", not just "values that apply to interactions between people".

---

1. [http://lesswrong.com/lw/sn/interpersonal\\_morality/](http://lesswrong.com/lw/sn/interpersonal_morality/)

2. Page 864, 'The Bedrock of Fairness'.

3. Page 864, 'The Bedrock of Fairness'.

The ordinary usage would have that jumping on a trampoline is not “morality”, it is just some selfish fun. On the other hand, giving someone else a turn to jump on the trampoline, is more akin to “morality” in common usage; and if you say “Everyone should take turns!” that’s definitely “morality”.

But the thing-I-want-to-talk-about includes the Fun Theory of a single person jumping on a trampoline.

Think of what a disaster it would be if all fun were removed from human civilization! So I consider it quite *right* to jump on a trampoline. Even if one would not say, in ordinary conversation, “I am jumping on that trampoline because I have a moral obligation to do so.” (Indeed, that sounds rather dull, and not at all fun, which is another important element of my “morality”.)

Alas, I do get the impression that in a standard academic discussion, one would use the term “morality” to refer to the sum-of-all-valu(ation rule)s that I am talking about. If there’s a standard alternative term in moral philosophy then do *please* let me know.

If there’s a better term than “morality” for the sum of all values and valuation rules, then this would free up “morality” for interpersonal values, which is closer to the common usage.

Some years ago, I was pondering what to say to the old cynical argument: If two monkeys want the same banana, in the end one will have it, and the other will cry morality.<sup>4</sup> I think the particular context was about whether the word “rights”, as in the context of “individual rights”, meant anything. It had just been vehemently asserted (on the Extropians mailing list, I think) that this concept was meaningless and ought to be tossed out the window.

Suppose there are two people, a Mugger and a Muggee. The Mugger wants to take the Muggee’s wallet. The Muggee doesn’t

---

4. [http://www.baen.com/library/0671318578/0671318578\\_33.htm](http://www.baen.com/library/0671318578/0671318578_33.htm)

want to give it to him. A cynic might say: “There is nothing more to say than this; they disagree. What use is it for the Muggee to claim that he has an *individual\_right* to keep his wallet? The Mugger will just claim that he has an *individual\_right* to take the wallet.”

Now today I might introduce the notion of a 1-place versus 2-place function<sup>5</sup>, and reply to the cynic, “Either they do not mean the same thing by *individual\_right*, or at least one of them is very mistaken about what their common morality implies.” At most one of these people is controlled by a good approximation of what I name when I say “morality”, and the other one is definitely not.

But the cynic might just say again, “So what? That’s what *you* say. The Mugger could just say the opposite. What meaning is there in such claims? What difference does it make?”

So I came up with this reply: “Suppose that *I* happen along this mugging. I will decide to side with the Muggee, not the Mugger, because I have the notion that the Mugger is interfering with the Muggee’s *individual\_right* to keep his wallet, rather than the Muggee interfering with the Mugger’s *individual\_right* to take it. And if a fourth person comes along, and must decide whether to allow my intervention, or alternatively stop me from treating on the Mugger’s *individual\_right* to take the wallet, then they are likely to side with the idea that I can intervene against the Mugger, in support of the Muggee.”

Now this does not work as a metaethics; it does not work to define the word *should*. If you fell backward in time, to an era when no one on Earth thought that slavery was wrong, you *should* still help slaves escape their owners. Indeed, the era when such an act was done in heroic defiance of society and the law, was not so very long ago.

But to defend the notion of *individual\_rights* against the charge of *meaninglessness*, the notion of third-party interventions and fourth-party allowances of those interventions, seems to me to coherently cash out *what is asserted* when we assert that an *individual\_right* exists. To assert that someone has a *right* to keep their wallet, is to assert that third parties *should* help them keep it, and that fourth parties *should* applaud those who thus help.

---

5. Page 849, ‘2-Place and 1-Place Words’.

This perspective does make a good deal of what is said about individual\_rights into nonsense<sup>6</sup>. “Everyone has a right to be free from starvation!” Um, who are you talking to? Nature? Perhaps you mean, “If you’re starving, and someone else has a hamburger, I’ll help you take it.” If so, you should say so clearly. (See also The Death of Common Sense<sup>7</sup>.)

So that is a notion of individual\_rights, but what does it have to do with the more general question of interpersonal morality?

The notion is that you can construct interpersonal morality out of individual morality. Just as, in this particular example, I constructed the notion of *what is* asserted by talking about an individual\_right, by making it an assertion about whether third parties should decide, for themselves, to interfere; and whether fourth parties should, individually, decide to applaud the interference.

Why go to such lengths to define things in individual terms? Some people might say: “To assert the existence of a right, is to say what *society* should do.”

But societies don’t always agree on things. And then you, as an individual, will have to decide what’s *right* for *you* to do, in that case.

“But individuals don’t always agree within themselves, either,” you say. “They have emotional conflicts.”

Well... you *could* say that and it would sound wise<sup>8</sup>. But generally speaking, neurologically intact humans will end up *doing some particular thing*. As opposed to flopping around on the floor as their limbs twitch in different directions under the temporary control of different personalities. Contrast to a government or a corporation<sup>9</sup>.

A human brain is a coherently adapted system<sup>10</sup> whose parts have been together optimized for a common criterion of fitness (more or less). A group is not functionally optimized as a group<sup>11</sup>. (You can verify this very quickly by looking at the sex ratios in a maternity hospital<sup>12</sup>.) *Individuals* may be optimized to do well out of

6. Page 180, ‘Applause Lights’.

7. <http://www.amazon.com/Death-Common-Sense-Suffocating-America/dp/0446672289>

8. Page 465, ‘How to Seem (and Be) Deep’.

9. [http://lesswrong.com/lw/l6/no\\_evolutions\\_for\\_corporations\\_or\\_nanodevices/](http://lesswrong.com/lw/l6/no_evolutions_for_corporations_or_nanodevices/)

10. [http://lesswrong.com/lw/l1/evolutionary\\_psychology/](http://lesswrong.com/lw/l1/evolutionary_psychology/)

11. [http://lesswrong.com/lw/kw/the\\_tragedy\\_of\\_group\\_selectionism/](http://lesswrong.com/lw/kw/the_tragedy_of_group_selectionism/)

12. [http://lesswrong.com/lw/l5/evolving\\_to\\_extinction/](http://lesswrong.com/lw/l5/evolving_to_extinction/)

their collective interaction—but that is quite a different selection pressure, the adaptations for which do not always produce group agreement! So if you want to look at a coherent decision system, it really is a good idea to look at one human, rather than a bureaucracy.

I myself am one person—admittedly with a long trail of human history behind me that makes me what I am, maybe more than any thoughts I ever thought myself. But still, at the end of the day, *I* am writing this blog post; it is not the negotiated output of a consortium. It is quite easy for me to imagine being faced, as an individual, with a case where the local group does not agree within itself—and in such a case I must decide, as an individual, what is *right*. In general I must decide what is right! If I go along with the group that does not absolve me of responsibility. If there are any countries that think differently, they can write their own blog posts.

This perspective, which does not exhibit undefined behavior in the event of a group disagreement, is one reason why I tend to treat interpersonal morality as a special case of individual morality, and not the other way around.

Now, with that said, interpersonal morality is a *highly distinguishable* special case of morality.

As humans, we don't just hunt in groups, we argue in groups. We've probably been arguing linguistically in adaptive political contexts for long enough—hundreds of thousands of years, maybe millions—to have adapted specifically to that selection pressure.

So it shouldn't be all that surprising if we have moral intuitions, like *fairness*, that apply specifically to the morality of groups.

One of these intuitions seems to be *universalizability*.

If Dennis just strides around saying, “I want the whole pie! Give me the whole pie! What's *fair* is for me to get the whole pie! Not you, me!” then that's not going to persuade anyone else in the tribe. Dennis has not managed to frame his desires in a form which enable them to leap from one mind to another. His desires will not take wings and become interpersonal. He is not likely to leave many offspring.

Now, the evolution of interpersonal moral intuitions, is a topic which (he said, smiling grimly) deserves its own blog post. And its

own academic subfield. (Anything out there besides *The Evolutionary Origins of Morality*? It seemed to me very basic.)

But I do think it worth noting that, rather than trying to manipulate 2-person and 3-person and 7-person interactions, some of our moral instincts seem to have made the leap to N-person interactions. We just think about *general moral arguments*. As though the values that leap from mind to mind, take on a life of their own and become something that you can reason about. To the extent that everyone in your environment *does* share some values, this will work as adaptive cognition. This creates moral intuitions that are not just *interpersonal* but *transpersonal*.

Transpersonal moral intuitions are not necessarily false-to-fact, so long as you don't expect your arguments cast in "universal" terms to sway a rock. There really is such a thing as the psychological unity of humankind<sup>13</sup>. Read a morality tale from an entirely different culture; I bet you can figure out what it's *trying* to argue *for*, even if you don't agree with it.

The problem arises when you try to apply the universalizability instinct to say, "If this argument could not persuade an UnFriendly AI that tries to maximize the number of paperclips in the universe, then it must not be a good argument."

There are No Universally Compelling Arguments<sup>14</sup>, so if you try to apply the universalizability instinct universally, you end up with no morality. Not even universalizability; the paperclip maximizer has no intuition of universalizability. It just chooses that action which leads to a future containing the maximum number of paperclips.

There are some things you just can't have a moral conversation with. There is not that within them that could respond to your arguments. You should think twice and maybe three times before ever saying this about one of your fellow humans—but a paperclip maximizer is another matter. You'll just have to override your moral instinct to regard anything labeled a "mind" as a little floating ghost-in-the-machine, with a hidden core of perfect emptiness, which could surely be persuaded to reject its mistaken source code if you just came up with the right argument. If you're going to pre-

---

13. [http://lesswrong.com/lw/rl/the\\_psychological\\_unity\\_of\\_humankind/](http://lesswrong.com/lw/rl/the_psychological_unity_of_humankind/)

14. Page 844, 'No Universally Compelling Arguments'.

serve universalizability as an intuition, you can try extending it to all humans; but you can't extend it to rocks or chatbots, nor even powerful optimization processes like evolutions<sup>15</sup> or paperclip maximizers.

The question of how much *in-principle agreement* would exist among human beings about the transpersonal portion of their values, given perfect knowledge of the facts and perhaps a much wider search of the argument space, is not a matter on which we can get much evidence by observing the prevalence of moral agreement and disagreement in today's world. Any disagreement might be something that the truth could destroy<sup>16</sup>—dependent on a different view of how the world is<sup>17</sup>, or maybe just dependent on having not yet heard the right argument. It is also possible that knowing more could dispel illusions of moral agreement<sup>18</sup>, not just produce new accords.

But does that question really make much difference in day-to-day moral reasoning, if you're *not* trying to build a Friendly AI?

---

15. [http://lesswrong.com/lw/kt/evolutions\\_are\\_stupid\\_but\\_work\\_anyway/](http://lesswrong.com/lw/kt/evolutions_are_stupid_but_work_anyway/)

16. Page 625, 'Feeling Rational'.

17. [http://lesswrong.com/lw/l4/terminal\\_values\\_and\\_instrumental\\_values/](http://lesswrong.com/lw/l4/terminal_values_and_instrumental_values/)

18. <http://www.somethinghappens.net/d/20080402.html>

## 28. Morality as Fixed Computation<sup>1</sup>

**Followup to:** The Meaning of Right<sup>2</sup>

Toby Ord commented<sup>3</sup>:

Eliezer, I've just reread your article and was wondering if this is a good quick summary of your position (leaving apart how you got to it):

'I should X' means that I would attempt to X were I fully informed.

Toby's a pro<sup>4</sup>, so if he didn't get it, I'd better try again. Let me try a different tack of explanation—one closer to the historical way that I arrived at my own position.

Suppose you build an AI, and—leaving aside that AI goal systems cannot be built around English statements<sup>5</sup>, and all such descriptions are only dreams—you try to infuse the AI with the action-determining principle, "Do what I want."

And suppose you get the AI design close *enough*—it doesn't just end up tiling the universe with paperclips, cheesecake or tiny molecular copies of satisfied programmers—that its utility function actually assigns utilities as follows, to the world-states we would describe in English as:

```
<Programmer weakly desires 'X',  
 quantity 20 of X exists>: +20  
<Programmer strongly desires 'Y',  
 quantity 20 of X exists>: 0  
<Programmer weakly desires 'X',  
 quantity 30 of Y exists>: 0  
<Programmer strongly desires 'Y',  
 quantity 30 of Y exists>: +60
```

---

1. [http://lesswrong.com/lw/sw/morality\\_as\\_fixed\\_computation/](http://lesswrong.com/lw/sw/morality_as_fixed_computation/)

2. Page 972, 'The Meaning of Right'.

3. [http://lesswrong.com/lw/sm/the\\_meaning\\_of\\_right/toz](http://lesswrong.com/lw/sm/the_meaning_of_right/toz)

4. <http://www.amirrorclear.net/academic/index.html>

5. [http://lesswrong.com/lw/lb/the\\_hidden\\_complexity\\_of\\_wishes/](http://lesswrong.com/lw/lb/the_hidden_complexity_of_wishes/)

You perceive, of course, that this destroys the world.

...since if the programmer initially weakly wants ‘X’ and X is hard to obtain, the AI will modify the programmer to strongly want ‘Y’, which is easy to create, and then bring about lots of Y. Y might be, say, iron atoms—those are highly stable.

Can you patch this problem? No. As a general rule, it is not possible to patch flawed Friendly AI designs.<sup>6</sup>

If you try to bound the utility function, or make the AI not care about how *much* the programmer wants things, the AI still has a motive (as an *expected* utility maximizer) to make the programmer want something that can be obtained with a very high degree of certainty.

If you try to make it so that the AI can’t modify the programmer, then the AI can’t talk to the programmer (talking to someone modifies them).

If you try to rule out a specific class of ways the AI could modify the programmer, the AI has a motive to superintelligently seek out loopholes and ways to modify the programmer indirectly.

As a general rule, it is not possible to patch flawed FAI designs.

We, ourselves, do not imagine the future and judge<sup>7</sup>, that any future in which our brains want something, and that thing exists, is a good future. If we did think this way, we would say: “Yay! Go ahead and modify us to strongly want something cheap!” But we do *not* say this, which means that this AI design is *fundamentally* flawed: it will choose things very unlike what we would choose; it will judge desirability very differently from how we judge it. This core disharmony cannot be patched by ruling out a handful of specific failure modes.<sup>8</sup>

There’s also a duality between Friendly AI problems and moral philosophy problems—though you’ve got to structure that duality in exactly the right way. So if you prefer, the core problem is that the AI will choose in a way very unlike the structure of what is, y’know, actually *right*—never mind the way we choose. Isn’t the whole point of this problem, that merely *wanting* something doesn’t *make it right*?

---

6. [http://lesswrong.com/lw/ld/the\\_hidden\\_complexity\\_of\\_wishes/](http://lesswrong.com/lw/ld/the_hidden_complexity_of_wishes/)

7. Page 948, ‘Math is Subjunctively Objective’.

8. [http://lesswrong.com/lw/ld/the\\_hidden\\_complexity\\_of\\_wishes/](http://lesswrong.com/lw/ld/the_hidden_complexity_of_wishes/)

So this is the paradoxical-seeming issue which I have analogized to the difference between:

A calculator that, when you press ‘2’, ‘+’, and ‘3’, tries to compute:

“What is  $2 + 3$ ? ”

A calculator that, when you press ‘2’, ‘+’, and ‘3’, tries to compute:

“What does this calculator output when you press ‘2’, ‘+’, and ‘3’? ”

The Type 1 calculator, as it were, *wants* to output 5.

The Type 2 “calculator” could return any result; and in the act of returning that result, it *becomes* the correct answer to the question that was internally asked.

We ourselves are like unto the Type 1 calculator. But the putative AI is being built as though it were to reflect the Type 2 calculator.

Now imagine that the Type 1 calculator is trying to build an AI, only the Type 1 calculator doesn’t *know* its own question. The calculator continually asks the question by its very nature, it was born to ask that question, created already in motion<sup>9</sup> around that question—but the calculator has no insight into its own transistors; it cannot print out the question, which is extremely complicated<sup>10</sup> and has no simple approximation<sup>11</sup>.

So the calculator wants to build an AI (it’s a pretty smart calculator, it just doesn’t have access to its own transistors) and have the AI give the right answer. Only the calculator can’t print out the question. So the calculator wants to have the AI look at the calculator, where the question is written, and answer the question that the AI will discover implicit in those transistors. But this cannot be done by the cheap shortcut of a utility function that says “All X: <calculator asks ‘X’?, answer X>: utility 1; else: utility 0” because that actually mirrors the utility function of a Type 2 calculator, not a Type 1 calculator.

---

9. Page 861, ‘Created Already In Motion’.

10. [http://lesswrong.com/lw/l3/thou\\_art\\_godsbatter/](http://lesswrong.com/lw/l3/thou_art_godsbatter/)

11. [http://lesswrong.com/lw/lq/fake\\_utility\\_functions/](http://lesswrong.com/lw/lq/fake_utility_functions/)

This gets us into FAI issues that I am not going into (some of which I'm still working out myself).

However, when you back out of the details of FAI design, and swap back to the perspective of moral philosophy, then *what we were just talking about* was the dual of the moral issue: “But if what’s ‘right’ is a mere preference, then anything that anyone wants is ‘right’.”

Now I did argue against that particular concept in some detail, in *The Meaning of Right*<sup>12</sup>, so I am not going to repeat all that...

But the key notion is the idea that what we name by ‘right’ is a *fixed* question, or perhaps a *fixed framework*. We can encounter moral arguments that modify our terminal values, and even encounter moral arguments that modify what we count as a moral argument; nonetheless, it all grows out of a particular starting point. We do not experience ourselves as embodying the question “What will I decide to do?” which would be a Type 2 calculator; anything we decided would thereby become right. We experience ourselves as asking the embodied question: “What will save my friends, and my people, from getting hurt? How can we all have more fun? ...” where the “...” is around a thousand other things.

So ‘I should X’ does not mean that I would attempt to X were I fully informed.

‘I should X’ means that X answers the question, “What will save my people? How can we all have more fun? How can we get more control over our own lives? What’s the funniest jokes we can tell? ...”

And I may not *know* what this question *is*, actually; I may not be able to print out my current guess nor my surrounding framework; but I know, as all non-moral-relativists instinctively know, that the question *surely* is not just “How can I do whatever I want?”

When these two formulations begin to seem as entirely distinct as “snow” and snow, then you shall have created distinct buckets<sup>13</sup> for the quotation and the referent<sup>14</sup>.

---

<sup>12</sup>. Page 972, ‘The Meaning of Right’.

<sup>13</sup>. Page 253, ‘Fallacies of Compression’.

<sup>14</sup>. Page 692, ‘The Quotation is not the Referent’.

**Added:** This was posted automatically and the front page got screwed up somehow. I have no idea how. It is now fixed and should make sense.

## 29. Inseparably Right; or, Joy in the Merely Good<sup>1</sup>

**Followup to:** The Meaning of Right<sup>2</sup>

I fear that in my drive for full explanation, I may have obscured the punchline from my theory of metaethics<sup>3</sup>. Here then is an attempted rephrase:

There is no pure ghostly essence of goodness apart from things like truth, happiness and sentient life.

What do you value? At a guess, you value the life of your friends and your family and your Significant Other and yourself, all in different ways. You would probably say that you value human life in general, and I would take your word for it<sup>4</sup>, though Robin Hanson might ask how you've acted on this supposed preference. If you're reading this blog you probably attach some value to truth for the sake of truth<sup>5</sup>. If you've ever learned to play a musical instrument, or paint a picture, or if you've ever solved a math problem for the fun of it, then you probably attach real value to good art. You value your freedom, the control that you possess over your own life; and if you've ever really helped someone you probably enjoyed it. You might not think of playing a video game as a great sacrifice of dutiful morality, but I for one would not wish to see the joy of complex challenge perish from the universe. You may not think of telling jokes as a matter of interpersonal morality<sup>6</sup>, but I would consider the human sense of humor as part of the gift we give to tomorrow<sup>7</sup>.

And you value many more things<sup>8</sup> than these.

Your brain assesses these things I have said, or others, or more, depending on the specific event, and finally affixes a little internal representational label that we recognize and call "good".

There's no way you can detach the little label from what it stands for, and still make ontological or moral sense.

---

1. [http://lesswrong.com/lw/sx/inseparably\\_right\\_or\\_joy\\_in\\_the\\_merely\\_good/](http://lesswrong.com/lw/sx/inseparably_right_or_joy_in_the_merely_good/)

2. Page 972, 'The Meaning of Right'.

3. Page 972, 'The Meaning of Right'.

4. [http://lesswrong.com/lw/b7/selfdeception\\_hypocrisy\\_or\\_akrasia/](http://lesswrong.com/lw/b7/selfdeception_hypocrisy_or_akrasia/)

5. Page 644, 'Something to Protect'.

6. Page 992, 'Interpersonal Morality'.

7. Page 927, 'The Gift We Give To Tomorrow'.

8. [http://lesswrong.com/lw/l3/thou\\_art\\_godsbatter/](http://lesswrong.com/lw/l3/thou_art_godsbatter/)

Why might the little ‘good’ label *seem* detachable? A number of reasons<sup>9</sup>.

Mainly, that’s just how your mind is structured—the labels it attaches internally seem like extra, floating, ontological properties<sup>10</sup>.

And there’s no *one* value that determines whether a complicated event is good or not—and no five values, either. No matter what rule you try to describe, there’s always something left over, some counterexample. Since no single value defines goodness, this can make it seem like all of them together couldn’t define goodness<sup>11</sup>. But when you add them up all together, there is nothing else left.

If there’s no detachable property of goodness, what does this mean?

It means that the question, “Okay, but what makes happiness or self-determination, *good*? ” is either very quickly answered, or else malformed.

The concept of a “utility function” or “optimization criterion” is detachable when talking about optimization processes. Natural selection, for example, optimizes for inclusive genetic fitness. But there are possible minds that implement any utility function<sup>12</sup>, so you don’t get any advice there about what you *should* do. You can’t ask about utility apart from any utility function.

When you ask “But which utility function *should* I use?” the word *should* is something inseparable from the dynamic that labels a choice “should”—inseparable from the reasons like “Because I can save more lives that way.”

Every time you say *should*, it includes an implicit criterion of choice; there is no should-ness that can be abstracted away from any criterion.

There is no separable right-ness that you could abstract from pulling a child off the train tracks, and attach to some other act.

Your values can change in response to arguments<sup>13</sup>; you have metamorals as well as morals. So it probably does make sense to think of an idealized good, or idealized right, that you would assign

9. Page 972, ‘The Meaning of Right’.

10. Page 685, ‘Mind Projection Fallacy’.

11. [http://lesswrong.com/lw/rc/the\\_ultimate\\_source/](http://lesswrong.com/lw/rc/the_ultimate_source/)

12. [http://lesswrong.com/lw/rm/the\\_design\\_space\\_of\\_mindsingeneral/](http://lesswrong.com/lw/rm/the_design_space_of_mindsingeneral/)

13. Page 924, ‘Whither Moral Progress?’.

if you could think of all possible arguments. Arguments may even convince you to change your criteria of what counts as a persuasive argument. Even so, when you consider the total trajectory arising out of that *entire framework*, that *moral frame of reference*, there is no separable property of justification-ness, apart from any particular criterion of justification; no final answer apart from a starting question.

I sometimes say that morality is “created already in motion<sup>14</sup>”.

There is no perfect argument that persuades the ideal philosopher of perfect emptiness to attach a perfectly abstract label of ‘good’. The notion of the perfectly abstract label is incoherent, which is why people chase it round and round in circles. What would distinguish a perfectly empty label of ‘good’ from a perfectly empty label of ‘bad’? How would you tell which was which?

But since every supposed criterion of goodness that we describe, turns out to be wrong, or incomplete, or changes the next time we hear a moral argument, it’s easy to see why someone might think that ‘goodness’ was a thing apart from any criterion at all.

Humans have a cognitive architecture that easily misleads us into conceiving of goodness as something that can be detached from any criterion.

This conception turns out to be incoherent. Very sad. I too was hoping for a perfectly abstract argument; it appealed to my universalizing<sup>15</sup> instinct. But...

But the question then becomes: is that little fillip of human psychology, more important than everything else? Is it more important than the happiness of your family, your friends, your mate, your extended tribe, and yourself? If your universalizing instinct is frustrated, is that worth abandoning life? If you represented rightness wrongly, do pictures stop being beautiful and maths stop being elegant? Is that one tiny mistake worth forsaking the gift we could give to tomorrow<sup>16</sup>? Is it even really worth all that much in the way of existential angst?

Or will you just say “Oops” and go back to life, to truth, fun, art, freedom, challenge, humor, moral arguments, and all those oth-

<sup>14</sup>. Page 861, ‘Created Already In Motion’.

<sup>15</sup>. Page 992, ‘Interpersonal Morality’.

<sup>16</sup>. Page 927, ‘The Gift We Give To Tomorrow’.

er things that in their sum and in their reflective trajectory, are the entire and only meaning of the word ‘right’?

Here is the strange habit of thought I mean to convey: Don’t look to some surprising<sup>17</sup> unusual<sup>18</sup> twist of logic for your justification. Look to the living child, successfully dragged off the train tracks. There you will find your justification. What ever should be more important than that?

I could dress that up in computational metaethics and FAI theory<sup>19</sup>—which indeed is whence the notion first came to me—but when I translated it all back into human-talk, that is what it turned out to say.

If we cannot take joy in things that are merely good, our lives shall be empty indeed.

---

17. <http://www.singinst.org/blog/2007/06/16/transhumanism-as-simplified-humanism/>

18. [http://lesswrong.com/lw/qz/living\\_in\\_many\\_worlds/](http://lesswrong.com/lw/qz/living_in_many_worlds/)

19. Page 999, ‘Morality as Fixed Computation’.

## 30. Sorting Pebbles Into Correct Heaps<sup>1</sup>

### Followup to: Anthropomorphic Optimism<sup>2</sup>

Once upon a time there was a strange little species—that might have been biological, or might have been synthetic, and perhaps were only a dream—whose passion was sorting pebbles into correct heaps.

They couldn't tell you *why* some heaps were correct, and some incorrect. But all of them agreed that the most important thing in the world was to create correct heaps, and scatter incorrect ones.

Why the Pebblesorting People cared so much, is lost to this history—maybe a Fisherian runaway sexual selection<sup>3</sup>, started by sheer accident a million years ago? Or maybe a strange work of sentient art, created by more powerful minds and abandoned?

But it mattered so drastically to them, this sorting of pebbles, that all the Pebblesorting philosophers said in unison that pebble-heap-sorting was the very meaning of their lives: and held that the only justified reason to eat was to sort pebbles, the only justified reason to mate was to sort pebbles, the only justified reason to participate in their world economy was to efficiently sort pebbles.

The Pebblesorting People all agreed on that, but they didn't always agree on which heaps were correct or incorrect.

In the early days of Pebblesorting civilization, the heaps they made were mostly small, with counts like 23 or 29; they couldn't tell if larger heaps were correct or not. Three millennia ago, the Great Leader Biko made a heap of 91 pebbles and proclaimed it correct, and his legions of admiring followers made more heaps likewise. But over a handful of centuries, as the power of the Bikonians faded, an intuition began to accumulate among the smartest and most educated that a heap of 91 pebbles was incorrect. Until finally they came to know what they had done: and they scattered all the heaps of 91 pebbles. Not without flashes of regret, for some of those heaps were great works of art, but incorrect. They even scattered Biko's original heap, made of 91 precious gemstones each of a different type and color.

---

1. [http://lesswrong.com/lw/sy/sorting\\_pebbles\\_into\\_correct\\_heaps/](http://lesswrong.com/lw/sy/sorting_pebbles_into_correct_heaps/)

2. [http://lesswrong.com/lw/st/anthropomorphic\\_optimism/](http://lesswrong.com/lw/st/anthropomorphic_optimism/)

3. [http://en.wikipedia.org/wiki/Fisherian\\_runaway](http://en.wikipedia.org/wiki/Fisherian_runaway)

And no civilization since has seriously doubted that a heap of 91 is incorrect.

Today, in these wiser times, the size of the heaps that Pebble-sorters dare attempt, has grown very much larger—which all agree would be a most great and excellent thing, if only they could ensure the heaps were really *correct*. Wars have been fought between countries that disagree on which heaps are correct: the Pebblesorters will never forget the Great War of 1957, fought between Y'ha-nthlei and Y'not'ha-nthlei, over heaps of size 1957. That war, which saw the first use of nuclear weapons on the Pebblesorting Planet, finally ended when the Y'not'ha-nthleian philosopher At'gra'len'ley exhibited a heap of 103 pebbles and a heap of 19 pebbles side-by-side. So persuasive was this argument that even Y'not'ha-nthlei reluctantly conceded that it was best to stop building heaps of 1957 pebbles, at least for the time being.

Since the Great War of 1957, countries have been reluctant to openly endorse or condemn heaps of large size, since this leads so easily to war. Indeed, some Pebblesorting philosophers—who seem to take a tangible delight in shocking others with their cynicism—have entirely denied the existence of pebble-sorting *progress*; they suggest that opinions about pebbles have simply been a random walk over time, with no coherence to them, the illusion of progress created by condemning all dissimilar pasts as incorrect. The philosophers point to the disagreement over pebbles of large size, as proof that there is nothing that makes a heap of size 91 really *incorrect*—that it was simply fashionable to build such heaps at one point in time, and then at another point, fashionable to condemn them. “But... 13!” carries no truck with them; for to regard “13!” as a persuasive counterargument, is only another convention, they say. The Heap Relativists claim that their philosophy may help prevent future disasters like the Great War of 1957, but it is widely considered to be a philosophy of despair.

Now the question of what makes a heap correct or incorrect, has taken on new urgency; for the Pebblesorters may shortly embark on the creation of self-improving Artificial Intelligences. The Heap Relativists have warned against this project: They say that AIs, not being of the species *Pebblesorter sapiens*, may form their own culture with entirely different ideas of which heaps are correct or incorrect. “They could decide that heaps of 8 pebbles are correct,”

say the Heap Relativists, “and while ultimately they’d be no righter or wronger than us, still, *our* civilization says we shouldn’t build such heaps. It is not in our interest to create AI, unless all the computers have bombs strapped to them, so that even if the AI thinks a heap of 8 pebbles is correct, we can force it to build heaps of 7 pebbles instead. Otherwise, KABOOM!”

But this, to most Pebblesorters, seems absurd. Surely a sufficiently powerful AI—especially the “superintelligence” some transpebblesorterists go on about—would be able to see *at a glance* which heaps were correct or incorrect! The thought of something with a brain the size of a planet, thinking that a heap of 8 pebbles was correct, is just too absurd to be worth talking about.

Indeed, it is an utterly futile project to constrain how a superintelligence sorts pebbles into heaps. Suppose that Great Leader Biko had been able, in his primitive era, to construct a self-improving AI; and he had built it as an expected utility maximizer whose utility function told it to create as many heaps as possible of size 91. Surely, when this AI improved itself far enough, and became smart enough, then it would see at a glance that this utility function was incorrect; and, having the ability to modify its own source code, it would *rewrite its utility function* to value more reasonable heap sizes, like 101 or 103.

And certainly not heaps of size 8. That would just be *stupid*. Any mind that stupid is too dumb to be a threat.

Reassured by such common sense, the Pebblesorters pour full speed ahead on their project to throw together lots of algorithms at random on big computers until some kind of intelligence emerges. The whole history of civilization has shown that richer, smarter, better educated civilizations are likely to agree about heaps that their ancestors once disputed. Sure, there are then larger heaps to argue about—but the further technology has advanced, the larger the heaps that have been agreed upon and constructed.

Indeed, intelligence itself has always correlated with making correct heaps—the nearest evolutionary cousins to the Pebblesorters, the Pebpanzees, make heaps of only size 2 or 3, and occasionally stupid heaps like 9. And other, even less intelligent creatures, like fish, make no heaps at all.

Smarter minds equal smarter heaps. Why would that trend break?

## 31. Moral Error and Moral Disagreement<sup>1</sup>

**Followup to:** Inseparably Right<sup>2</sup>, Sorting Pebbles Into Correct Heaps<sup>3</sup>

Richard Chappell, a pro<sup>4</sup>, writes<sup>5</sup>:

“When Bob says “Abortion is wrong”, and Sally says, “No it isn’t”, they are disagreeing with each other.

I don’t see how Eliezer can accommodate this. On his account, what Bob asserted is true iff abortion is prohibited by the morality\_Bob norms. How can Sally disagree? There’s no disputing (we may suppose) that abortion is indeed prohibited by morality\_Bob...

Since there is moral disagreement, whatever Eliezer purports to be analysing here, it is not morality.”

The phenomena of moral disagreement, moral error, and moral progress, on terminal values<sup>6</sup>, are *the* primary drivers behind my metaethics<sup>7</sup>. Think of how simple Friendly AI would be if there were no moral disagreements, moral errors, or moral progress!

Richard claims, “There’s no disputing (we may suppose) that abortion is indeed prohibited by morality\_Bob.”

We may *not* suppose, and there *is* disputing. Bob does not have direct, unmediated, veridical access to the output of his own morality.

I tried to describe morality as a “computation<sup>8</sup>”. In retrospect, I don’t think this is functioning as the Word of Power<sup>9</sup> that I thought I was emitting<sup>10</sup>.

---

1. [http://lesswrong.com/lw/sz/moral\\_error\\_and\\_moral\\_disagreement/](http://lesswrong.com/lw/sz/moral_error_and_moral_disagreement/)

2. Page 1004, ‘Inseparably Right; or, Joy in the Merely Good’.

3. Page 1008, ‘Sorting Pebbles Into Correct Heaps’.

4. <http://www.philosophyetc.net/2007/01/is-normativity-just-semantics.html>

5. [http://lesswrong.com/lw/sx/inseparably\\_right\\_or\\_joy\\_in\\_the\\_merely\\_good/mde](http://lesswrong.com/lw/sx/inseparably_right_or_joy_in_the_merely_good/mde)

6. [http://lesswrong.com/lw/l4/terminal\\_values\\_and\\_instrumental\\_values/](http://lesswrong.com/lw/l4/terminal_values_and_instrumental_values/)

7. Page 972, ‘The Meaning of Right’.

8. Page 999, ‘Morality as Fixed Computation’.

9. [http://lesswrong.com/lw/kg/expecting\\_short\\_inferential\\_distances/](http://lesswrong.com/lw/kg/expecting_short_inferential_distances/)

10. [http://lesswrong.com/lw/ke/illusion\\_of\\_transparency\\_wby\\_no\\_one\\_understands/](http://lesswrong.com/lw/ke/illusion_of_transparency_wby_no_one_understands/)

Let us read, for “computation”, “idealized abstract dynamic”—maybe that will be a more comfortable label to apply to morality.

Even so, I would have thought it obvious that computations may be the subjects of mystery and error. Maybe it’s not as obvious outside computer science?

Disagreement has two prerequisites: the possibility of agreement and the possibility of error. For two people to agree on something, there must be something they are agreeing *about*, a referent held in common. And it must be possible for an “error” to take place, a conflict between “P” in the map and not-P in the territory. Where these two prerequisites are present, Sally can say to Bob: “That thing we were just both talking about—you are in error about it.”

Richard’s objection would seem in the first place to rule out the possibility of moral error, from which he derives the impossibility of moral agreement.

So: does my metaethics rule out moral error? Is there no disputing that abortion is indeed prohibited by morality\_Bob?

This is such a strange idea that I find myself wondering what the heck Richard could be thinking. My best guess is that Richard, perhaps having not read all the posts in this sequence, is taking my notion of morality\_Bob to refer to a *flat, static list of valuations explicitly asserted by Bob*. “Abortion is wrong” would be on Bob’s list, and there would be no disputing that.

But on the contrary, I conceive of morality\_Bob as something that *unfolds* into Bob’s morality—like the way one can describe in 6 states and 2 symbols<sup>ii</sup> a Turing machine that will write  $4.640 \times 10^{1439}$  1s to its tape before halting.

So morality\_Bob refers to a compact folded specification, and not a flat list of outputs. But still, how could Bob be wrong about the output of his own morality?

In manifold obvious and non-obvious ways:

Bob could be empirically mistaken about the state of fetuses, perhaps believing fetuses to be aware of the outside world. (Cor-

---

<sup>ii</sup>. [http://en.wikipedia.org/wiki/Busy\\_beaver#current\\_6-state.2C\\_2-symbol\\_best\\_contender](http://en.wikipedia.org/wiki/Busy_beaver#current_6-state.2C_2-symbol_best_contender)

recting this might change Bob's instrumental values but not terminal values<sup>12</sup>.)

Bob could have formed his beliefs about what constituted "personhood" in the presence of confusion<sup>13</sup> about the nature of consciousness, so that if Bob were fully informed about consciousness, Bob would not have been tempted to talk about "the beginning of life" or "the human kind" in order to define personhood. (This changes Bob's expressed terminal values; afterward he will state different general rules about what sort of physical things are ends in themselves.)

So those are the obvious moral errors—instrumental errors driven by empirical mistakes; and erroneous generalizations about terminal values, driven by failure to consider moral arguments that are valid but hard to find in the search space.

Then there are less obvious sources of moral error: Bob could have a list of mind-influencing considerations that he considers morally valid, and a list of other mind-influencing considerations that Bob considers morally invalid. Maybe Bob was raised a Christian and now considers that cultural influence to be invalid. But, unknown to Bob, when he weighs up his values for and against abortion, the influence of his Christian upbringing comes in and distorts his summing of value-weights. So Bob believes that the output of his current validated moral beliefs is to prohibit abortion, but actually this is a leftover of his childhood and not the output of those beliefs at all.

(Note that Robin Hanson and I seem to disagree, in a case like this, as to exactly what degree we should take Bob's word about what his morals are<sup>14</sup>.)

Or Bob could believe that the word of God determines moral truth and that God has prohibited abortion in the Bible. Then Bob is making metaethical mistakes, causing his mind to malfunction in a highly general way, and add moral generalizations to his belief pool, which he would not do if veridical knowledge of the universe destroyed his current and incoherent metaethics.

Now let us turn to the disagreement between Sally and Bob.

12. [http://lesswrong.com/lw/l4/terminal\\_values\\_and\\_instrumental\\_values/](http://lesswrong.com/lw/l4/terminal_values_and_instrumental_values/)

13. Page 679, 'Wrong Questions'.

14. [http://lesswrong.com/lw/b7/selfdeception\\_by\\_pocrysis\\_or\\_akrasia/](http://lesswrong.com/lw/b7/selfdeception_by_pocrysis_or_akrasia/)

You could suggest that Sally is saying to Bob, “Abortion is allowed by morality\_Bob”, but that seems a bit oversimplified; it is not psychologically or morally realistic.

If Sally and Bob were unrealistically sophisticated, they might describe their dispute as follows:

Bob: “Abortion is wrong.”

Sally: “Do you think that this is something of which most humans ought to be persuadable?”

Bob: “Yes, I do. Do you think abortion is right?”

Sally: “Yes, I do. And I don’t think that’s because I’m a psychopath by common human standards. I think most humans would come to agree with me, if they knew the facts I knew, and heard the same moral arguments I’ve heard.”

Bob: “I think, then, that we must have a moral disagreement: since we both believe ourselves to be a shared moral frame of reference on this issue, and yet our moral intuitions say different things to us.”

Sally: “Well, it is not *logically necessary* that we have a genuine disagreement. We might be mistaken in believing ourselves to mean the same thing by the words *right* and *wrong*, since neither of us can introspectively report our own moral reference frames or unfold them fully.”

Bob: “But if the meaning is similar up to the third decimal place, or sufficiently similar in some respects that it ought to be delivering similar answers on *this particular* issue, then, even if our moralities are not in principle *identical*, I would not hesitate to invoke the intuitions for transpersonal morality<sup>15</sup>.”

---

<sup>15</sup>. Page 992, ‘Interpersonal Morality’.

Sally: “I agree. Until proven otherwise, I am inclined to talk about this question as if it is the same question unto us.”

Bob: “So I say ‘Abortion is wrong’ without further qualification or specialization on what *wrong* means unto me.”

Sally: “And I think that abortion is right. We have a disagreement, then, and at least one of us must be mistaken.”

Bob: “Unless we’re *actually* choosing differently *because* of in-principle unresolvable differences in our moral frame of reference, as if one of us were a paperclip maximizer. In that case, we would be mutually mistaken in our belief that when we talk about doing what is right, we mean the same thing by *right*. We would agree that we have a disagreement, but we would both be wrong.”

Now, this is not exactly what most people are explicitly thinking when they engage in a moral dispute—but it is how I would cash out and naturalize their intuitions about transpersonal morality.

Richard also says, “Since there is moral disagreement...” This seems like a prime case of what I call naive<sup>16</sup> philosophical<sup>17</sup> realism<sup>18</sup>—the belief that philosophical intuitions are direct unmediated veridical passports to philosophical truth.

It so happens that I agree that there *is* such a thing as moral disagreement. Tomorrow I will endeavor to justify, in fuller detail, how this statement can possibly make sense in a reductionistic natural universe. So I am not disputing this particular *proposition*. But I note, in passing, that Richard cannot justifiably assert the existence of moral disagreement as an *irrefutable premise* for discussion, though he could consider it as an *apparent datum*. You cannot take as irrefutable premises, things that you have not explained exactly; for then what is it that is certain to be true?

<sup>16</sup>. Page 944, ‘Can Counterfactuals Be True?’.

<sup>17</sup>. [http://lesswrong.com/lw/my/the\\_allais\\_paradox/](http://lesswrong.com/lw/my/the_allais_paradox/)

<sup>18</sup>. Page 751, ‘Hand vs. Fingers’.

I cannot help but note the resemblance to Richard's assumption that "there's no disputing" that abortion is indeed prohibited by morality\_Bob—the assumption that Bob has direct veridical unmediated access to the final unfolded output of his own morality.

Perhaps Richard means that we *could* suppose that abortion is indeed prohibited by morality\_Bob, and allowed by morality\_Sally, there being at least two possible minds for whom this would be true. Then the two minds might be mistaken about believing themselves to disagree. Actually they would simply be directed by different algorithms.

You cannot have a disagreement about which algorithm *should* direct your actions, without first having the same meaning of *should*—and no matter how you try to phrase this in terms of "what ought to direct your actions" or "right actions" or "correct heaps of pebbles"<sup>19</sup>, in the end you will be left with the empirical fact that it is possible to construct<sup>20</sup> minds directed by any coherent utility function.

When a paperclip maximizer and a pencil maximizer do different things, they are not *disagreeing* about anything, they are just different optimization processes. You cannot detach should-ness<sup>21</sup> from any specific criterion of should-ness and be left with a pure empty should-ness that the paperclip maximizer and pencil maximizer can be said to *disagree* about—unless you cover "disagreement" to include differences where two agents have nothing to say to each other.

But this would be an extreme position to take with respect to your fellow humans, and I recommend against doing so. Even a psychopath would still be in a common moral reference frame with you, if, fully informed, they would decide to take a pill that would make them non-psychopaths. If you told me that my ability to care about other people was neurologically damaged, and you offered me a pill to fix it, I would take it. Now, perhaps some psychopaths would not be persuadable in-principle to take the pill that would, by our standards, "fix" them. But I note the possibility to emphasize what an extreme statement it is to say of someone:

---

19. Page 1008, 'Sorting Pebbles Into Correct Heaps'.

20. [http://lesswrong.com/lw/rm/the\\_design\\_space\\_of\\_mindsingeneral/](http://lesswrong.com/lw/rm/the_design_space_of_mindsingeneral/)

21. Page 1004, 'Inseparably Right; or, Joy in the Merely Good'.

“We have nothing to argue about, we are only different optimization processes.”

That should be reserved for paperclip maximizers, not used against humans whose arguments you don’t like.

## 32. Abstracted Idealized Dynamics<sup>1</sup>

**Followup to:** Morality as Fixed Computation<sup>2</sup>

I keep trying to describe morality as a “computation<sup>3</sup>”, but people don’t stand up and say “Aha!”

Pondering the surprising inferential distances<sup>4</sup> that seem to be at work here, it occurs to me that when I say “computation”, some of my listeners may not hear the Word of Power<sup>5</sup> that I thought I was emitting<sup>6</sup>; but, rather, may think of some complicated boring unimportant thing like Microsoft Word.

Maybe I should have said that morality is an *abstracted idealized dynamic*. This might not have meant anything to start with, but at least it wouldn’t sound like I was describing Microsoft Word.

How, oh how, am I to describe the awesome import of this concept, “computation”?

Perhaps I can display the inner nature of computation, in its most general form, by showing how that inner nature manifests in something that seems very unlike Microsoft Word—namely, morality.

Consider certain features we might wish to ascribe to that-which-we-call “morality”, or “should” or “right” or “good”:

- It seems that we sometimes think about morality in our armchairs, without further peeking at the state of the outside world, and arrive at some previously unknown conclusion.

Someone sees a slave being whipped, and it doesn’t occur to them right away that slavery is wrong. But they go home and think about it, and imagine themselves in the slave’s place, and finally think, “No.”

Can you think of anywhere else that something like this happens?

Suppose I tell you that I am making a rectangle of pebbles. You look at the rectangle, and count 19 pebbles on one side and 103 dots

---

1. [http://lesswrong.com/lw/to/abstracted\\_idealized\\_dynamics/](http://lesswrong.com/lw/to/abstracted_idealized_dynamics/)

2. Page 999, ‘Morality as Fixed Computation’.

3. Page 999, ‘Morality as Fixed Computation’.

4. [http://lesswrong.com/lw/kg/expecting\\_short\\_inferential\\_distances/](http://lesswrong.com/lw/kg/expecting_short_inferential_distances/)

5. [http://lesswrong.com/lw/kg/expecting\\_short\\_inferential\\_distances/](http://lesswrong.com/lw/kg/expecting_short_inferential_distances/)

6. [http://lesswrong.com/lw/ke/illusion\\_of\\_transparency\\_wby\\_no\\_one\\_understands/](http://lesswrong.com/lw/ke/illusion_of_transparency_wby_no_one_understands/)

pebbles on the other side. You don't know right away how many pebbles there are. But you go home to your living room, and draw the blinds, and sit in your armchair and think; and without further looking at the physical array, you come to the conclusion that the rectangle contains 1957 pebbles.

Now, I'm not going to say the word "computation". But it seems like that—which—is "morality" should have the property of *latent development of answers*—that you may not know right away, everything that you have sufficient in-principle information to know. All the ingredients are present, but it takes additional time to bake the pie.

You can specify a Turing machine of 6 states and 2 symbols<sup>7</sup> that unfolds into a string of  $4.6 \times 10^{1439}$  is after  $2.5 \times 10^{2879}$  steps. A machine I could describe aloud in ten seconds, runs longer and produces a larger state than the whole observed universe to date.

When you distinguish between the program *description* and the program's *executing state*, between the process specification and the final outcome, between the question and the answer, you can see why even certainty about a program description does not imply human certainty about the executing program's outcome. See also Artificial Addition<sup>8</sup> on the difference between a compact specification versus a flat list of outputs.

Morality, likewise, is something that unfolds, through arguments, through discovery, through thinking; from a bounded set of intuitions and beliefs that animate our initial states, to a potentially much larger set of specific moral judgments we may have to make over the course of our lifetimes.

- When two human beings both think about the same moral question, even in a case where they both start out uncertain of the answer, it is not unknown for them to come to the same conclusion. It seems to happen more often than chance alone would allow—though the biased focus of reporting and memory<sup>9</sup> is on the shouting and the arguments. And this is so, even if both humans remain in their armchairs and do not peek out the living-room blinds while thinking.

---

7. [http://en.wikipedia.org/wiki/Busy\\_beaver#current\\_6-state.2C\\_2-symbol\\_best\\_contender](http://en.wikipedia.org/wiki/Busy_beaver#current_6-state.2C_2-symbol_best_contender)

8. [http://lesswrong.com/lw/ly/artificial\\_addition/](http://lesswrong.com/lw/ly/artificial_addition/)

9. <http://lesswrong.com/lw/j5/availability/>

Where else does this happen? It happens when trying to guess the number of pebbles in a rectangle of sides 19 and 103. Now this does not prove by Greek analogy<sup>10</sup> that morality is multiplication. If A has property X and B has property X it does not follow that A is B. But it seems that morality ought to have the property of *expected agreement about unknown latent answers*, which, please note, generally implies that *similar questions are being asked in different places*.

This is part of what is conveyed by the Word of Power, “computation”: the notion of similar questions being asked in different places and having similar answers. Or as we might say in the business, the same computation can have multiple instantiations.

If we know the structure of calculator 1 and calculator 2, we can decide that they are “asking the same question” and that we ought to see the “same result” flashing on the screen of calculator 1 and calculator 2 after pressing the Enter key. We decide this in advance of seeing the actual results, which is what makes the concept of “computation” predictively useful.

And in fact, we can make this deduction even without knowing the exact circuit diagrams of calculators 1 and 2, so long as we’re told that the circuit diagrams are the same.

And then when we see the result “1957” flash on the screen of calculator 1, we know that the same “1957” can be expected to flash on calculator 2, and we even expect to count up 1957 pebbles in the array of 19 by 103.

A hundred calculators, performing the same multiplication in a hundred different ways, can be expected to arrive at the same answer—and this is not a vacuous expectation adduced after seeing similar answers. We can form the expectation in *advance* of seeing the actual answer.

Now this does not show that morality is in fact a little electronic calculator. But it highlights the notion of something that *factors out* of different physical phenomena in different physical places, even phenomena as physically different as a calculator and an array of pebbles—a common answer to a common question. (Where is this factored-out thing? Is there an Ideal Multiplication Table written on a stone tablet somewhere outside the universe? But we are not concerned with that for now.)

---

<sup>10</sup>. [http://lesswrong.com/lw/rj/surface\\_analogies\\_and\\_deep\\_causes/](http://lesswrong.com/lw/rj/surface_analogies_and_deep_causes/)

Seeing that one calculator outputs “1957”, we infer that *the answer*—the *abstracted answer*—is 1957; and from there we make our predictions of what to see on all the other calculator screens, and what to see in the array of pebbles.

So that-which-we-name-morality seems to have the further properties of *agreement about developed latent answers*, which we may as well think of in terms of *abstract answers*; and note that such agreement is unlikely in the absence of *similar questions*.

- We sometimes look back on our own past moral judgments, and say “Oops!” E.g., “Oops! Maybe in retrospect I shouldn’t have killed all those guys when I was a teenager.”

So by now it seems easy to extend the analogy, and say: “Well, maybe a cosmic ray hits one of the transistors in the calculator and it says ‘1959’ instead of 1957—that’s an error.”

But this notion of “error”, like the notion of “computation” itself, is more subtle than it appears.

Calculator Q says ‘1959’ and calculator X says ‘1957’. Who says that calculator Q is wrong, and calculator X is right? Why not say that calculator X is wrong and calculator Q is right? Why not just say, “the results are different”?

“Well,” you say, drawing on your store of common sense, “if it was just those two calculators, I wouldn’t know for sure which was right. But here I’ve got nine other calculators that all say ‘1957’, so it certainly seems *probable* that 1957 is the correct answer.”

What’s this business about “correct”? Why not just say “different”?

“Because if I have to predict the outcome of any other calculators that compute  $19 \times 103$ , or the number of pebbles in a  $19 \times 103$  array, I’ll predict 1957—or whatever observable outcome corresponds to the abstract number 1957.”

So perhaps  $19 \times 103 = 1957$  only most of the time. Why call the answer 1957 the *correct* one, rather than the mere fad among calculators, the majority vote?

If I’ve got a hundred calculators, all of them rather error-prone—say a 10% probability of error—then there is no *one* calculator I can point to and say, “This is the standard!” I might pick a calculator that would happen, on this occasion, to vote with ten other calculators rather than ninety other calculators. This is

why I have to *idealize* the answer, to talk about this *ethereal* thing that is not associated with any particular physical process known to me—not even arithmetic done in my own head, which can also be “incorrect”.

It is this ethereal process, this idealized question, to which we compare the results of any one particular calculator, and say that the result was “right” or “wrong”.

But how can we obtain information about this perfect and unphysical answer, when all that we can ever observe, are merely physical phenomena? Even doing “mental” arithmetic just tells you about the result in your own, merely physical brain<sup>ii</sup>.

“Well,” you say, “the pragmatic answer is that we can obtain extremely strong evidence by looking at the results of a hundred calculators, even if they are only 90% likely to be correct on any one occasion.”

But wait: When do electrons or quarks or magnetic fields ever make an “error”? If no individual particle can be mistaken, how can any collection of particles be mistaken? The concept of an “error”, though humans may take it for granted, is hardly something that would be mentioned in a fully reductionist view of the universe.

Really, what happens is that we have a certain model in mind of the calculator—the model that we looked over and said, “This implements  $19 * 103$ ”—and then other physical events caused the calculator to depart from this model, so that the final outcome, while physically lawful, did not correlate with that mysterious abstract thing, and the other physical calculators, in the way we had in mind. Given our mistaken beliefs about the physical process of the first calculator, we would look at its output ‘1959’, and make mistaken predictions about the other calculators (which do still hew to the model we have in mind).

So “incorrect” cashes out, naturalistically, as “physically departed from the model that I had of it” or “physically departed from the idealized question that I had in mind”. A calculator struck by a cosmic ray, is not ‘wrong’ in any physical sense, not an unlawful event in the universe; but the outcome is not the answer to the question you had in mind, the question that you believed empirically-falsely the calculator would correspond to.

---

ii. [http://lesswrong.com/lw/k2/a\\_priori/](http://lesswrong.com/lw/k2/a_priori/)

The calculator's "incorrect" answer, one might say, is an answer to a different question than the one you had in mind—it is an empirical fact about the calculator that it implements a different computation.

- The 'right' act or the 'should' option sometimes seem to depend on the state of the physical world. For example, should you cut the red wire or the green wire to disarm the bomb?

Suppose I show you a long straight line of pebbles, and ask you, "How many pebbles would I have, if I had a rectangular array of six lines like this one?" You start to count, but only get up to 8 when I suddenly blindfold you.

Now you are not completely ignorant of the answer to this question. You know, for example, that the result will be even, and that it will be greater than 48. But you can't answer the question until you know how many pebbles were in the original line.

But mark this about the question: It wasn't a question about anything you could directly see in the world, at that instant. There was not in fact a rectangular array of pebbles, six on a side. You *could* perhaps lay out an array of such pebbles and count the results—but then there are more complicated computations that we could run on the unknown length of a line of pebbles. For example, we could treat the line length as the start of a Goodstein sequence<sup>12</sup>, and ask whether the sequence halts. To physically play out this sequence would require many more pebbles than exist in the universe. Does it make sense to ask if the Goodstein sequence which starts with the length of this line of pebbles, "would halt"? Does it make sense to talk about *the answer*, in a case like this?

I'd say yes, personally.

But meditate upon the etherealness of the answer—that we talk about idealized abstract processes that never really happen; that we talk about what *would* happen if the law of the Goodstein sequence came into effect upon this line of pebbles, even though the law of the Goodstein sequence will never physically come into effect.

It is the same sort of etherealness that accompanies the notion of a proposition that  $19 * 103 = 1957$  which factors out of any particular physical calculator and is not identified with the result of any particular physical calculator.

---

<sup>12.</sup> [http://en.wikipedia.org/wiki/Goodstein%27s\\_theorem](http://en.wikipedia.org/wiki/Goodstein%27s_theorem)

Only now that etherealness has been mixed with physical things; we talk about the effect of an ethereal operation on a physical thing. We talk about what would happen if we ran the Goodstein process on *the number of pebbles in this line here*, which we have not counted—we do not know exactly how many pebbles there are. There is no tiny little XML tag upon the pebbles that says “Goodstein halts”, but we still think—or at least I still think—that it makes sense to say of the pebbles that they have the property of their Goodstein sequence terminating.

So computations can be, as it were, idealized abstract *dynamics*—idealized abstract applications of idealized abstract laws, iterated over an imaginary causal-time that could go on for quite a number of steps (as Goodstein sequences often do).

So when we wonder, “*Should* I cut the red wire or the green wire?”, we are not multiplying or simulating the Goodstein process, in particular. But we are wondering about something that is not physically immanent in the red wires or the green wires themselves; there is no little XML tag on the green wire, saying, “This is the wire that *should* be cut.”

We may not know which wire defuses the bomb, but say, “Whichever wire does in fact defuse the bomb, that is the wire that *should* be cut.”

Still, there are no little XML tags on the wires, and we may not even have any way to look inside the bomb—we may just have to guess, in real life.

So if we try to cash out this notion of a definite wire that *should* be cut, it’s going to come out as...

...some rule that would tell us which wire to cut, if we knew the exact state of the physical world...

...which is to say, some kind of idealized abstract process into which we feed the state of the world as an input, and get back out, “cut the green wire” or “cut the red wire”...

...which is to say, the output of a computation that would take the world as an input.

- And finally I note that from the twin phenomena of *moral agreement* and *moral error*, we can construct the notion of *moral disagreement*.

This adds nothing to our understanding of “computation” as a Word of Power, but it’s helpful in putting the pieces together.

Let’s say that Bob and Sally are talking about an abstracted idealized dynamic they call “Enamuh”.

Bob says “The output of Enamuh is ‘Cut the blue wire’,” and Sally says “The output of Enamuh is ‘Cut the brown wire’.”

Now there are several non-exclusive possibilities:

Either Bob or Sally could have committed an error in applying the rules of Enamuh—they could have done the equivalent of mis-multiplying known inputs.

Either Bob or Sally could be mistaken about some empirical state of affairs upon which Enamuh depends—the wiring of the bomb.

Bob and Sally could be talking about different things when they talk about Enamuh, in which case both of them are committing an error when they refer to Enamuh\_Bob and Enamuh\_Sally by the same name. (However, if Enamuh\_Bob and Enamuh\_Sally differ in the sixth decimal place in a fashion that doesn’t change the output about which wire gets cut, Bob and Sally can quite legitimately gloss the difference.)

Or if Enamuh itself is defined by some other abstracted idealized dynamic, a Meta-Enamuh whose output is Enamuh, then either Bob or Sally could be mistaken about Meta-Enamuh in any of the same ways they could be mistaken about Enamuh. (But in the case of morality, we have an abstracted idealized dynamic that includes a specification of how it, itself, changes. Morality is *self-renormalizing*—it is not a guess at the product of some different and outside source.)

To sum up:

- Morality, like computation, involves *latent development of answers*;
- Morality, like computation, permits *expected agreement of unknown latent answers*;
- Morality, like computation, reasons about *abstract results apart from any particular physical implementation*;
- Morality, like computation, *unfolds from bounded initial state into something potentially much larger*;

- Morality, like computation, can be viewed as *an idealized dynamic that would operate on the true state of the physical world*—permitting us to speak about idealized answers of which we are physically uncertain;
- Morality, like computation, lets us to speak of such unphysical stuff as “error”, by *comparing a physical outcome to an abstract outcome*—presumably in a case where there was previously reason to believe or desire that the physical process was isomorphic to the abstract process, yet this was not actually the case.

And so with all that said, I hope that the word “computation” has come to convey something other than Microsoft Word.

## 33. “Arbitrary”<sup>1</sup>

**Followup to:** Inseparably Right; or, Joy in the Merely Good<sup>2</sup>, Sorting Pebbles Into Correct Heaps<sup>3</sup>

One of the experiences of following the Way is that, from time to time, you notice a new word that you have been using without really understanding. And you say: “What does this word, ‘X’, really mean?”

Perhaps ‘X’ is ‘error’, for example. And those who have not yet realized the importance of this aspect of the Way, may reply: “Huh? What do you mean? Everyone knows what an ‘error’ is; it’s when you get something wrong, when you make a mistake.” And you reply, “But those are only synonyms; what can the term ‘error’ mean<sup>4</sup> in a universe where particles only ever do what they do?”

It’s not meant to be a rhetorical question; you’re meant to go out and answer it. One of the primary tools for doing so is Rationalist’s Taboo<sup>5</sup>, when you try to speak without using the word or its synonyms—to replace the symbol with the substance<sup>6</sup>.

So I ask you therefore, what is this word “arbitrary”? Is a rock arbitrary? A leaf? A human?

How about sorting pebbles into prime-numbered heaps? How about maximizing inclusive genetic fitness? How about dragging a child off the train tracks?

How can I tell exactly which things are arbitrary, and which not, in this universe where particles only ever do what they do? Can you tell me exactly what property is being discriminated, without using the word “arbitrary” or any direct synonyms? Can you open up the box of “arbitrary”, this label that your mind assigns to some things and not others, and tell me what kind of algorithm is at work here?

Having pondered this issue myself, I offer to you the following proposal:

---

1. <http://lesswrong.com/lw/t1/arbitrary/>

2. Page 1004, ‘Inseparably Right; or, Joy in the Merely Good’.

3. Page 1008, ‘Sorting Pebbles Into Correct Heaps’.

4. Page 1019, ‘Abstracted Idealized Dynamics’.

5. Page 245, ‘Taboo Your Words’.

6. Page 248, ‘Replace the Symbol with the Substance’.

A piece of cognitive content feels “arbitrary” if it is the kind of cognitive content that we expect to come with attached justifications, and those justifications are not present in our mind.

You’ll note that I’ve performed the standard operation for guaranteeing that a potentially confusing question has a real answer:<sup>7</sup> I substituted the question, “How does my brain label things ‘arbitrary?’” for “What is this mysterious property of arbitrariness?” This is not necessarily a sleight-of-hand, since to explain something is not the same as explaining it away<sup>8</sup>.

In this case, for nearly all everyday purposes, I would make free to proceed from “arbitrary” to arbitrary. If someone says to me, “I believe that the probability of finding life on Mars is  $6.203 \times 10^{-23}$  to four significant digits,” I would make free to respond, “That sounds like a rather arbitrary number,” not “My brain has attached the subjective arbitrariness-label to its representation of the number in your belief.”

So as it turned out in this case, having answered the question “What is ‘arbitrary?’” turns out not to affect the way I use the word ‘arbitrary’; I am just more aware of what the arbitrariness-sensation indicates. I am aware that when I say, “ $6.203 \times 10^{-23}$  sounds like an arbitrary number”, I am indicating that I would expect some justification for assigning that particular number, and I haven’t heard it. This also explains why the precision is important—why I would question that particular number, but not someone saying “Less than 1%”. In the latter case, I have some idea what might justify such a statement; but giving a very precise figure implies that you have some kind of information I don’t know about, either that or you’re being silly.

“Ah,” you say, “but what do you mean by ‘justification’? Haven’t you failed to make any progress, and just passed the recursive buck<sup>9</sup> to another black box<sup>10</sup>? ”

Actually, no; I told you that “arbitrariness” was a sensation produced by the *absence* of an *expected* X. Even if I don’t tell you

---

<sup>7</sup>. Page 682, ‘Righting a Wrong Question’.

<sup>8</sup>. Page 703, ‘Explaining vs. Explaining Away’.

<sup>9</sup>. [http://lesswrong.com/lw/rd/passing\\_the\\_recursive\\_buck/](http://lesswrong.com/lw/rd/passing_the_recursive_buck/)

<sup>10</sup>. Page 154, ‘Mysterious Answers to Mysterious Questions’.

anything more about that X, you've learned something about the cognitive algorithm—opened up the original black box, and taken out two gears and a *smaller* black box.

But yes, it makes sense to continue onward to discuss this mysterious notion of “justification”.

Suppose I told you that “justification” is what tells you whether a belief is reasonable. Would this tell you anything? No, because there are no extra gears that have been factored out, just a direct invocation of “reasonable”-ness.

Okay, then suppose instead I tell you, “Your mind labels X as a justification for Y, whenever adding ‘X’ to the pool of cognitive content would result in ‘Y’ being added to the pool, or increasing the intensity associated with ‘Y’.” How about that?

“Enough of this buck-passing tomfoolery!” you may be tempted to cry. But wait; this really does factor out another couple of gears. We have the idea that different propositions, to the extent they are held, can create each other in the mind, or increase the felt level of intensity—credence for beliefs, desire for acts or goals. You may have already known this, more or less, but stating it aloud is still progress.

This may not provide much satisfaction to someone inquiring into morals. But then someone inquiring into morals may well do better to just think moral thoughts, rather than thinking about metaethics or reductionism.

On the other hand, if you were building a Friendly AI, and trying to explain to that FAI what a human being means by the term “justification”, then the statement I just issued might help the FAI narrow it down. With some additional guidance, the FAI might be able to figure out where to look, in an empirical model of a human, for representations of the sort of *specific* moral content that a human inquirer-into-morals would be interested in—what *specifically* counts or doesn't count as a justification, in the eyes of that human. And this being the case, you might not have to explain the specifics exactly correctly at system boot time; the FAI knows how to find out the rest on its own. My inquiries into metaethics are not directed toward the same purposes as those of standard philosophy.

Now of course you may reply, “Then the FAI finds out what the human *thinks* is a “justification”. But is that formulation of

‘justification’, really *justified?*“ But by this time, I hope, you can predict my answer to that sort of question, whether or not you agree. I answer that we have just witnessed a strange loop through the meta-level<sup>11</sup>, in which you use justification-as-justification to evaluate the quoted form of justification-as-cognitive-algorithm, which algorithm may, perhaps, happen to be your own, &c. And that the feeling of “justification” cannot be coherently detached<sup>12</sup> from the specific algorithm we use to decide justification in particular cases; that there is no pure empty essence of justification that will persuade any optimization process regardless of its algorithm, &c.

And the upshot is that differently structured minds may well label different propositions with their *analogues* of the internal label “arbitrary”—though only one of these labels is what *you mean* when you say “arbitrary”, so you and these other agents do not really have a disagreement.

---

11. Page 887, ‘Where Recursive Justification Hits Bottom’.

12. Page 1004, ‘Inseparably Right; or, Joy in the Merely Good’.

## 34. Is Fairness Arbitrary?<sup>1</sup>

**Followup to:** The Bedrock of Fairness<sup>2</sup>

In “The Bedrock of Fairness<sup>3</sup>”, Xannon, Yancy, and Zaire argue over how to split up a pie that they found in the woods. Yancy thinks that  $1/3$  each is fair; Zaire demands half; and Xannon tries to compromise.

Dividing a pie fairly isn’t as trivial a problem as it may sound. What if people have different preferences for crust, filling, and topping? Should they each start with a third, and trade voluntarily? But then they have conflicts of interest over how to divide the surplus utility generated by trading...

But I would say that “half for Zaire” surely isn’t fair.

I confess that I originally wrote Zaire as a foil—this is clearer in an earlier version of the dialog, where Zaire, named Dennis, demands the whole pie—and was surprised to find some of my readers taking Zaire’s claim seriously, perhaps because I had Zaire say “I’m hungry.”

Well, okay; I believe that when I write a dialogue, the reader has a right to their own interpretation. But I did intend that dialogue to illustrate a particular point:

You can argue about how to divide up the pie, or even argue how to argue about dividing up the pie, you can argue over what is fair... but there finally comes a point when you hit bedrock. If Dennis says, “No, the *fair* way to argue is that *I* get to dictate everything, and I now hereby dictate that I get the whole pie,” there’s nothing left to say but “Sorry, that’s just not what *fairness* is—you can try to take the pie and I can try to stop you, but you can’t convince that *that* is fair.”

A “fair division” is not the same as “a division that compels everyone to admit that the division is fair”. Dennis can always just refuse to agree, after all.

But more to the point, when you encounter a pie in the forest, in the company of friends, and you try to be *fair*, there’s a certain particular thing you’re trying to do—the term “fair” is not perfectly

---

1. [http://lesswrong.com/lw/t2/is\\_fairness\\_arbitrary/](http://lesswrong.com/lw/t2/is_fairness_arbitrary/)

2. Page 864, ‘The Bedrock of Fairness’.

3. Page 864, ‘The Bedrock of Fairness’.

empty, it cannot attach to just anything. Metaphorically speaking, “fair” is not a hypothesis equally compatible with any outcome<sup>4</sup>.

Fairness expresses notions of concern for the other agents who also want the pie; a goal to take their goals into account. It’s a separate question whether that concern is pure altruism, or not wanting to make them angry enough to fight. Fairness expresses notions of symmetry, equal treatment—which might be a terminal value unto you, or just an attempt to find a convenient meeting-point to avoid an outright battle.

Is it fair to take into account what *other* people think is “fair”, and not just what *you* think is “fair”?

The obvious reason to care what other people think is “fair”, is if they’re being moved by *similar considerations*, yet arriving at different conclusions. If you think that the Other’s word “fair” means what you think of as *fair*, and you think the Other is being honest about what they think, then you ought to pay attention just by way of fulfilling your *own* desire to be fair. It is like paying attention to an honest person who means the same thing you do by “multiplication”, who says that  $19 * 103$  might not be  $1947$ . The attention you pay to that suggestion, is not a favor to the other person; it is something you do if *you* want to get the multiplication right—*they’re* doing *you* a favor by correcting you.

Politics is more subject to bias than multiplication. And you might think that the Other’s reasoning is corrupted by self-interest, while yours is as pure as Antarctic snow. But to the extent that you credit the Other’s self-honesty, or doubt your own, you would do well to hear what the Other has to say—if *you* wish to be fair.

The second notion of why we might pay attention to what someone else thinks is “fair”, is more complicated: it is the notion of *applying fairness to its own quotation*, that is, fairly debating what is “fair”. In complicated politics you may have to negotiate a negotiating procedure. Surely it wouldn’t be fair if Dennis just got to say, “The fair resolution procedure is that I get to decide what’s fair.” So why should *you* get to just decide what’s fair, then?

Here the attention you pay to the other person’s beliefs about “fairness”, is a favor that *you* do to *them*, a concession that you expect to be met with a return concession.

---

4. Page 138, ‘Fake Explanations’.

But when you set out to fairly discuss what is “fair” (note the strange loop through the meta-levels), that doesn’t put *everything* up for grabs. A zeroth-order fair division of a pie doesn’t involve giving away the *whole* pie to Dennis—just giving identical portions to all. Even though Dennis wants the whole thing, and asks for the whole thing, the zeroth-order fair division only gives Dennis a symmetrical portion to everyone else’s. Similarly, a first-order fair attempt to resolve a dispute about what is “fair”, doesn’t involve conceding everything to the Other’s viewpoint without reciprocation. That wouldn’t be fair. Why give everything away to the Other, if you receive nothing in return? Why give Dennis the whole first-order pie?

On some level, then, there has to be a possible demand which would be too great—a demand exceeding what may be *fairly* requested of you. This is part of the content of fairness; it is part of what you are setting out to do, when you set out to be fair. Admittedly, one should not be too trigger-happy about saying “That’s too much!” We human beings tend to overestimate the concessions we have made, and underestimate the concessions that others have made to us; we tend to underadjust for the Other’s point of view... even so, if *nothing* is “too much”, then you’re not engaging in *fairness*.

Fairness might call on you to hear out what the Other has to say; fairness may call on you to exert an effort to really truly consider the Other’s point of view—but there is a limit to this, as there is a limit to all fair concessions. If all Dennis can say is “I want the whole pie!” over and over, there’s a limit to how long fairness requires you to ponder this argument.

You reach the bedrock of fairness at the point where, no matter who questions whether the division is fair, no matter who refuses to be persuaded, no matter who offers further objections, and regardless of your awareness that you yourself may be biased... Dennis still isn’t getting the whole pie. If there are others present who are also trying to be fair, and Dennis is not already dictator, they will probably back you rather than Dennis—this is one sign that you can trust the line you’ve drawn, that it really is time to say “Enough!”

If you and the others present get together and give Dennis 1/ $N$ th of the pie—or even if *you* happen to have the upper hand,

---

<sup>5</sup>. Page 887, ‘Where Recursive Justification Hits Bottom’.

and you unilaterally give Dennis and yourself and all others each  $1/N$ —then you are not being unfair on *any* level; there is no meta-level of fairness where Dennis gets the whole pie.

Now I'm sure there are some in the audience who will say, "You and perhaps some others, are *merely* doing things your way, rather than Dennis's." On the contrary: We are merely being fair. It so happens that this fairness is our way, as all acts must be *someone's* way to happen in the real universe. But what we are merely doing, happens to be, being *fair*. And there is no level on which it is unfair, because there is no level on which fairness requires unlimited unreciprocated surrender.

I don't believe in unchangeable bedrock—I believe in self-modifying bedrock<sup>6</sup>. But I do believe in bedrock, in the sense that everything has to start somewhere. It can be turtles all the way up, but not turtles all the way down.

You cannot define fairness *entirely* in terms of "That which everyone agrees is 'fair'." This isn't just nonterminating. It isn't just ill-defined if Dennis doesn't believe that 'fair' is "that which everyone agrees is 'fair'". It's actually *entirely empty*, like the English sentence "This sentence is true." Is that sentence true? Is it false? It is neither; it doesn't mean anything because it is entirely wrapped up in itself, with no tentacle of relation to reality. If you're going to argue what is fair, there has to be *something* you're arguing *about*, some structure that is baked into the question.

Which is to say that you can't turn "fairness" into an ideal label of pure emptiness, defined *only* by the mysterious compulsion of every possible agent to admit "This is what is 'fair'." Forget the case against universally compelling<sup>7</sup> arguments—just consider the definition itself: *It has absolutely no content, no external references*; it is not just underspecified, but *entirely unspecified*.

But as soon as you introduce any content into the label "fairness" that *isn't* phrased purely in terms of all possible minds applying the label, then you have a foundation on which to stand. It may be self-modifying bedrock, rather than immovable bedrock. But it is still a place to start. A place from which to say: "Regardless of

---

6. Page 887, 'Where Recursive Justification Hits Bottom'.

7. Page 844, 'No Universally Compelling Arguments'.

what Dennis says, giving him the whole pie *isn't fair*, because *fairness* is not defined entirely and only in terms of Dennis's agreement."

And you aren't being "arbitrary", either—though the intuitive meaning of that word has never seemed entirely well-specified to me; is a tree arbitrary, or a leaf? But it sounds like the accusation is of pulling some answer out of thin air—which you're *not* doing; you're giving the *fair* answer, not an answer pulled out of thin air. What about when you jump up a meta-level, and look at Dennis's wanting to do it one way, and your wanting a different resolution? Then it's still not arbitrary, because you aren't being *unfair* on that meta-level, either. The answer you pull out is not *merely* an arbitrary answer you invented, but a *fair* answer. You aren't *merely* doing it your way; the way that you are doing it, is the fair way.

You can ask "But why *should* you be fair?"—and that's a separate question, which we'll go into tomorrow. But giving Dennis  $1/N$ th, we can at least say, is not *merely and only arbitrary* from the perspective of fair-vs.-unfair. Even if Dennis keeps saying "It isn't fair!" and even if Dennis also disputes the 1st-order, 2nd-order, Nth-order meta-fairnesses. Giving N people each  $1/N$ th is nonetheless a *fair* sort of thing to do, and whether or not we *should* be fair is then a separate question.

## 35. The Bedrock of Morality: Arbitrary?<sup>1</sup>

**Followup to:** Is Fairness Arbitrary?<sup>2</sup>, Joy in the Merely Good<sup>3</sup>, Sorting Pebbles Into Correct Heaps<sup>4</sup>

Yesterday, I presented the idea that when only five people are present, having just stumbled across a pie in the woods (a naturally growing pie, that just popped out of the ground) then it is fair to give Dennis only 1/5th of this pie, even if Dennis persistently claims that it is fair for him to get the whole thing. Furthermore, it is meta-fair to follow such a symmetrical division procedure, even if Dennis insists that *he* ought to dictate the division procedure.

Fair, meta-fair, or meta-meta-fair, there is no level of fairness where you're obliged to concede everything to Dennis, without reciprocity or compensation, just because he demands it.

Which goes to say that fairness has a meaning beyond which "that which everyone can be convinced is 'fair'". This is an empty proposition, isomorphic to "Xyblz is that which everyone can be convinced is 'xyblz'". There must be some *specific* thing of which people are being convinced; and once you identify that thing, it has a meaning beyond agreements and convincing.

You're not introducing something *arbitrary*, something un-fair, in refusing to concede everything to Dennis. You are being fair, and meta-fair and meta-meta-fair. As far up as you go, there's no level that calls for unconditional surrender. The stars do not judge between you and Dennis—but it *is* baked into the very question that is asked, when you ask, "What is fair?" as opposed to "What is xyblz?"

Ah, but why *should* you be fair, rather than xyblz? Let us concede that Dennis cannot validly persuade us, on any level, that it is *fair* for him to dictate terms and give himself the whole pie; but perhaps he could argue whether we *should* be fair?

The hidden agenda of the whole discussion of fairness, of course, is that good-ness and right-ness and should-ness, ground out similarly to fairness.

---

1. [http://lesswrong.com/lw/t3/the\\_bedrock\\_of\\_morality\\_arbitrary/](http://lesswrong.com/lw/t3/the_bedrock_of_morality_arbitrary/)

2. Page 1032, 'Is Fairness Arbitrary?'.

3. Page 1004, 'Inseparably Right; or, Joy in the Merely Good'.

4. Page 1008, 'Sorting Pebbles Into Correct Heaps'.

Natural selection<sup>5</sup> optimizes for inclusive genetic fitness. This is not a disagreement<sup>6</sup> with humans about what is good. It is simply that natural selection does not *do* what is good: it optimizes for inclusive genetic fitness.

Well, since some optimization processes optimize for inclusive genetic fitness, instead of what is good, which *should* we do, ourselves?

I know my answer to this question. It has something to do with natural selection being a terribly wasteful and stupid<sup>7</sup> and inefficient process. It has something to do with elephants starving to death in their old age when they wear out their last set of teeth. It has something to do with natural selection never choosing a single act of mercy, of grace, even when it would cost its purpose nothing: not auto-anesthetizing a wounded and dying gazelle, when its pain no longer serves even the adaptive purpose that first created pain. Evolution had to happen sometime in the history of the universe, because that's the only way that intelligence could *first* come into being, without brains to make brains; but now that era is over, and good riddance.

But most of all—why on Earth *would* any human being think that one *ought* to optimize inclusive genetic fitness, rather than what is good? What is even the appeal of this, morally or otherwise? *At all?* I know people who claim<sup>8</sup> to think like this, and I wonder what wrong turn they made in their cognitive history, and I wonder how to get them to snap out of it.

When we take a step back from fairness, and ask if we *should* be fair, the answer may not always be yes. Maybe sometimes we should be merciful. But if you ask if it is *meta-fair* to be fair, the answer will generally be yes. Even if someone else wants you to be unfair in their favor, or claims to disagree about what is “fair”, it will still generally be meta-fair to be fair, even if you can't make the Other agree. By the same token, if you ask if we meta-should do what we should, rather than something else, the answer is yes. Even if some other agent or optimization process does not do what is right, that doesn't change what is meta-right.

---

5. [http://lesswrong.com/lw/kr/an\\_alien\\_god/](http://lesswrong.com/lw/kr/an_alien_god/)

6. Page 1019, ‘Abstracted Idealized Dynamics’.

7. [http://lesswrong.com/lw/kt/evolutions\\_are\\_stupid\\_but\\_work\\_anyway/](http://lesswrong.com/lw/kt/evolutions_are_stupid_but_work_anyway/)

8. Page 108, ‘Belief in Belief’.

And this is not “arbitrary” in the sense of rolling dice, not “arbitrary” in the sense that justification is expected and then not found. The accusations that I level against evolution are not *merely* pulled from a hat; they are expressions of morality as I understand it. They are merely moral, and there is nothing mere about that.

In “Arbitrary”<sup>9</sup> I finished by saying:

The upshot is that differently structured minds may well label different propositions with their *analogues* of the internal label “arbitrary”—though only one of these labels is what *you mean* when you say “arbitrary”, so you and these other agents do not really have a disagreement.

This was to help shake people loose of the idea that if any two possible minds can say or do different things, then it must all be arbitrary. Different minds may have different ideas of what’s “arbitrary”, so clearly this whole business of “arbitrariness” is arbitrary, and we should ignore it. After all, Sinned (the anti-Dennis) just always says “Morality isn’t arbitrary!” no matter how you try to persuade her otherwise, so clearly you’re just being arbitrary in saying that morality is arbitrary.

From the perspective of a human, saying that one should sort pebbles into prime-numbered heaps<sup>10</sup> is arbitrary—it’s the sort of act you’d expect to come with a justification attached, but there isn’t any justification.

From the perspective of a Pebblesorter, saying that one p-should scatter a heap of 38 pebbles into two heaps of 19 pebbles is not p-arbitrary at all—it’s the most p-important thing in the world, and fully p-justified by the intuitively obvious fact that a heap of 19 pebbles is p-correct and a heap of 38 pebbles is not.

So which perspective should we adopt? I answer that I see no reason at all why I should start sorting pebble-heaps. It strikes me as a completely pointless activity. Better to engage in art, or music, or science, or heck, better to connive political plots of terrifying dark elegance, than to sort pebbles into prime-numbered heaps. A galaxy transformed into pebbles and sorted into prime-numbered heaps would be just plain boring.

---

9. Page 1028, “Arbitrary”.

10. Page 1008, ‘Sorting Pebbles Into Correct Heaps’.

The Pebblesorters, of course, would only reason that music is p-pointless because it doesn't help you sort pebbles into heaps; the human activity of humor is not only p-pointless but just plain p-bizarre and p-incomprehensible; and most of all, the human vision of a galaxy in which agents are running around experiencing positive reinforcement *but not sorting any pebbles*, is a vision of an utterly p-arbitrary galaxy devoid of p-purpose. The Pebblesorters would gladly sacrifice their lives to create a P-Friendly AI that sorted the galaxy on their behalf; it would be the most p-profound statement they could make about the p-meaning of their lives.

So which of these two perspectives do I choose? The human one, of course; not because it is the human one, but because it is *right*. I do not know perfectly what is right, but neither can I plead entire ignorance<sup>11</sup>.

And the Pebblesorters, *who simply are not built to do what is right*, choose the Pebblesorting perspective: not merely because it is theirs, or because they think they can get away with being p-arbitrary, but because that is what is p-right.

And in fact, both we and the Pebblesorters can *agree* on all these points. We can agree that sorting pebbles into prime-numbered heaps is arbitrary and unjustified, but not p-arbitrary or p-unjustified; that it is the sort of thing an agent p-should do, but not the sort of thing an agent should do.

I fully expect that even if there is other life in the universe only a few trillions of lightyears away (I don't think it's local, or we would have seen it by now), that we humans are the only creatures for a long long way indeed who are built to do what is *right*. That may be a moral miracle<sup>12</sup>, but it is not a causal miracle.

There may be some other evolved races, a sizable fraction perhaps, maybe even a majority, who do some right things. Our executing adaptation<sup>13</sup> of compassion is not so far removed from the game theory that gave it birth; it might be a common adaptation. But laughter, I suspect, may be rarer by far than mercy. What would a galactic civilization be like, if it had sympathy, but never a moment of humor? A little more boring, perhaps, by our standards.

<sup>11</sup>. Page 935, 'Could Anything Be Right?'.

<sup>12</sup>. Page 927, 'The Gift We Give To Tomorrow'.

<sup>13</sup>. [http://lesswrong.com/lw/lo/adaptationexecuters\\_not\\_fitnessmaximizers/](http://lesswrong.com/lw/lo/adaptationexecuters_not_fitnessmaximizers/)

This humanity that we find ourselves in, is a great gift. It may not be a great p-gift, but who cares about p-gifts?

So I really must deny the charges of moral relativism: I don't think that human morality is arbitrary at all, and I would expect any logically omniscient reasoner to agree with me on that. We are better than the Pebblesorters, because we care about sentient lives, and the Pebblesorters don't. Just as the Pebblesorters are p-better than us, because they care about pebble heaps, and we don't. Human morality is p-arbitrary, but who cares? P-arbitrariness is arbitrary.

You've just got to avoid thinking that the words "better" and "p-better", or "moral" and "p-moral", are *talking about the same thing*—because then you might think that the Pebblesorters were coming to different conclusions than us about *the same thing*—and then you might be tempted to think that our own morals were arbitrary. Which, of course, they're not.

Yes, I really truly do believe that humanity is better than the Pebblesorters! I am not being sarcastic, I really do believe that. I am not playing games by redefining "good" or "arbitrary", I think I mean the same thing by those terms as everyone else. When you understand that I am genuinely sincere about that, you will understand my metaethics. I really *don't* consider myself a moral relativist—not even in the slightest!

## 36. You Provably Can't Trust Yourself<sup>1</sup>

**Followup to:** Where Recursive Justification Hits Bottom<sup>2</sup>, Löb's Theorem<sup>3</sup>

Peano Arithmetic *seems* pretty trustworthy. We've never found a case where Peano Arithmetic proves a theorem T, and yet T is false in the natural numbers. That is, we know of no case where  $\Box T$  ("T is provable in PA") and yet  $\neg T$  ("not T").

We also know of no case where first order logic is invalid: We know of no case where first-order logic produces *false conclusions* from *true premises*. (Whenever first-order statements H are true of a model, and we can syntactically deduce C from H, checking C against the model shows that C is also true.)

Combining these two observations, it seems like we should be able to get away with adding a rule to Peano Arithmetic that says:

All T:  $(\Box T \rightarrow T)$

But Löb's Theorem<sup>4</sup> seems to show that as soon as we do that, everything becomes provable. What went wrong? How can we do *worse* by adding a true premise to a trustworthy theory? Is the premise not true—does PA prove some theorems that are false? Is first-order logic not valid—does it sometimes prove false conclusions from true premises?

Actually, there's nothing wrong with reasoning from the axioms of Peano Arithmetic plus the axiom schema "Anything provable in Peano Arithmetic is true." But the result is a *different* system from PA, which we might call PA+I. PA+I does not reason from identical premises to PA; something new has been added. So we can evade Löb's Theorem because PA+I is not trusting *itself*—it is only trusting PA.

If you are not previously familiar with mathematical logic, you might be tempted to say, "Bah! Of course PA+I is trusting itself! PA+I just isn't willing to admit it! Peano Arithmetic *already* believes anything provable in Peano Arithmetic—it will *already* output anything provable in Peano Arithmetic as a theorem, *by definition!*"

---

1. [http://lesswrong.com/lw/t8/you\\_provably\\_cant\\_trust\\_yourself/](http://lesswrong.com/lw/t8/you_provably_cant_trust_yourself/)

2. Page 887, 'Where Recursive Justification Hits Bottom'.

3. [http://lesswrong.com/lw/t6/the\\_cartoon\\_guide\\_to\\_löbs\\_theorem/](http://lesswrong.com/lw/t6/the_cartoon_guide_to_löbs_theorem/)

4. [http://lesswrong.com/lw/t6/the\\_cartoon\\_guide\\_to\\_löbs\\_theorem/](http://lesswrong.com/lw/t6/the_cartoon_guide_to_löbs_theorem/)

How does moving to PA+I change anything, then? PA+I is just the same system as PA, and so by trusting PA, PA+I is really trusting itself. Maybe that dances around some obscure mathematical problem with direct self-reference, but it doesn't evade the charge of self-trust."

But PA+I and PA really are different systems; in PA+I it is possible to prove true statements about the natural numbers that are not provable in PA. If you're familiar with mathematical logic, you know this is because some nonstandard models of PA are ruled out in PA+I. Otherwise you'll have to take my word that Peano Arithmetic doesn't fully describe the natural numbers, and neither does PA+I, but PA+I characterizes the natural numbers slightly better than PA.

The deeper point is the *enormous* gap, the *tremendous* difference, between having a system just like PA except that it trusts PA, and a system just like PA except that it trusts *itself*.

If you have a system that trusts PA, that's no problem; we're pretty sure PA is trustworthy, so the system is reasoning from true premises. But if you have a system that looks like PA—having the standard axioms of PA—but also trusts *itself*, then it is trusting a self-trusting system, something for which there is no precedent. In the case of PA+I, PA+I is trusting PA which we're pretty sure is correct. In the case of Self-PA it is trusting Self-PA, which we've never seen before—it's never been tested, despite its *misleading* surface similarity<sup>5</sup> to PA. And indeed, Self-PA collapses via Löb's Theorem and proves everything—so I guess it *shouldn't* have trusted itself after all! All this isn't magic; I've got a nice Cartoon Guide<sup>6</sup> to how it happens, so there's no good excuse for not understanding what goes on here.

I<sup>7</sup> have<sup>8</sup> spoken<sup>9</sup> of the Type 1 calculator that asks “What is  $2 + 3$ ?” when the buttons “2”, “+”, and “3” are pressed; versus the Type 2 calculator that asks “What do I calculate when someone presses ‘2 + 3’?” The first calculator answers 5; the second calculator can truthfully answer anything, even 54.

5. [http://lesswrong.com/lw/rj/surface\\_analogies\\_and\\_deep\\_causes/](http://lesswrong.com/lw/rj/surface_analogies_and_deep_causes/)

6. [http://lesswrong.com/lw/t6/the\\_cartoon\\_guide\\_to\\_lafabs\\_theorem/](http://lesswrong.com/lw/t6/the_cartoon_guide_to_lafabs_theorem/)

7. Page 915, ‘Probability is Subjectively Objective’.

8. Page 999, ‘Morality as Fixed Computation’.

9. Page 1019, ‘Abstracted Idealized Dynamics’.

But this doesn't mean that all calculators that reason about calculators are flawed. If I build a third calculator that asks "What does the first calculator answer when I press '2 + 3?'", perhaps by calculating out the individual transistors, it too will answer 5. Perhaps this new, reflective calculator will even be able to answer some questions faster, by virtue of proving that some faster calculation is isomorphic to the first calculator.

$\text{PA}$  is the equivalent of the first calculator;  $\text{PA+I}$  is the equivalent of the third calculator; but Self- $\text{PA}$  is like unto the second calculator.

As soon as you start trusting yourself, you become unworthy of trust. You'll start believing *any* damn thing that you think, just because *you* thought it. This wisdom of the human condition is pleasingly analogous to a precise truth of mathematics.

Hence the saying: "Don't believe everything you think."

And the math also suggests, by analogy, how to do better<sup>10</sup>: Don't trust thoughts *because you think them*, but because they *obey specific trustworthy rules*.

$\text{PA}$  only starts believing something—metaphorically speaking—when it sees a specific proof, laid out in black and white. If you say to  $\text{PA}$ —even if you prove to  $\text{PA}$ —that  $\text{PA}$  will prove something,  $\text{PA}$  still won't believe you until it sees the *actual proof*. Now, this might seem to invite inefficiency, and  $\text{PA+I}$  will believe you—if you prove that  $\text{PA}$  will prove something, because  $\text{PA+I}$  trusts the specific, fixed framework of Peano Arithmetic; not *itself*.

As far as any human knows,  $\text{PA}$  does happen to be sound; which means that what  $\text{PA}$  proves is provable in  $\text{PA}$ ,  $\text{PA}$  *will* eventually prove and *will* eventually believe. Likewise, anything  $\text{PA+I}$  can prove that it proves, it will eventually prove and believe. It seems so tempting to just make  $\text{PA}$  trust *itself*—but then it becomes Self- $\text{PA}$  and implodes. Isn't that odd?  $\text{PA}$  believes everything it proves, but it doesn't believe "Everything I prove is true."  $\text{PA}$  trusts a fixed framework for how to prove things, and that framework doesn't happen to talk *about* trust in the framework.

You *can* have a system that trusts the  $\text{PA}$  framework *explicitly*, as well as implicitly: that is  $\text{PA+I}$ . But the new framework that  $\text{PA+I}$  *uses*, makes no *mention* of itself; and the specific proofs that  $\text{PA+I}$

---

<sup>10</sup> [http://lesswrong.com/lw/b8/tsuyoku\\_naritai\\_i\\_want\\_to\\_become\\_stronger/](http://lesswrong.com/lw/b8/tsuyoku_naritai_i_want_to_become_stronger/)

demands, make no mention of trusting PA+I, only PA. You might say that PA implicitly trusts PA, PA+I explicitly trusts PA, and Self-PA trusts itself.

For everything that you believe, you should always find yourself able to say, “I believe because of [specific argument in framework F]”, not “I believe because I believe”.

Of course, this gets us into the +I question of why you ought to trust or use framework F. Human beings, not being formal systems, are too reflective to get away with being *unable* to think about the problem. Got a superultimate framework U? Why trust U?

And worse: as far as I can tell, *using* induction is what leads me to *explicitly* say that induction seems to often work, and my *use* of Occam’s Razor is implicated in my explicit *endorsement* of Occam’s Razor. Despite my best efforts, I have been unable to prove that this is inconsistent, and I suspect it may be valid<sup>11</sup>.

But it does seem that the distinction between *using* a framework and *mentioning* it, or between *explicitly* trusting a fixed framework F and trusting *yourself*, is at least *important* to unraveling foundational tangles—even if Löb turns out not to apply directly.

Which gets me to the reason why I’m saying all this in the middle of a sequence about morality.

I’ve been pondering the unexpectedly large inferential distances<sup>12</sup> at work here—I thought I’d gotten all the prerequisites out of the way for explaining metaethics, but no. I’m no longer sure I’m even close. I tried to say that morality was a “computation”, and that failed; I tried to explain that “computation” meant “abstracted idealized dynamic”, but that didn’t work either. No matter how many different ways I tried to explain it, I couldn’t get across the distinction my metaethics drew between “do the right thing”, “do the human thing”, and “do my own thing”. And it occurs to me that my own background, coming into this, may have relied on having already drawn the distinction between PA, PA+I and Self-PA.

Coming to terms with metaethics, I am beginning to think, is all about *distinguishing between levels*. I first learned to do this *rigorously* back when I was getting to grips with mathematical logic, and discovering that you could prove complete absurdities<sup>13</sup>, if you

---

11. Page 887, ‘Where Recursive Justification Hits Bottom’.

12. [http://lesswrong.com/lw/kg/expecting\\_short\\_inferential\\_distances/](http://lesswrong.com/lw/kg/expecting_short_inferential_distances/)

lost track even once of the distinction between “believe particular PA proofs”, “believe PA is sound”, and “believe you yourself are sound”. If you believe any particular PA proof, that might sound pretty much the same as believing PA is sound in general; and if you use PA and only PA, then trusting PA (that is, being moved by arguments that follow it) sounds pretty much the same as believing that you yourself are sound. But after a bit of practice with the actual math—I did have to practice the actual math, not just read about it—my mind formed permanent distinct buckets and built walls around them to prevent the contents from slopping over.

Playing around with PA and its various conjugations, gave me the notion of what it meant to trust *arguments within a framework that defined justification*. It gave me practice keeping track of specific frameworks, and holding them distinct in my mind.

Perhaps that’s why I expected to communicate more sense than I actually succeeded in doing, when I tried to describe *right* as a framework of justification that involved being moved by particular, specific terminal values and moral arguments; analogous to an entity who is moved by encountering a specific proof from the allowed axioms of Peano Arithmetic. As opposed to a general license to do whatever you prefer, or a morally relativistic term like “utility function” that can eat the values of any given species, or a neurological framework contingent on particular facts about the human brain. You can make good use of such concepts, but I do not identify them with the substance of what is *right*.

Gödelian arguments are inescapable; you can always isolate the framework-of-trusted-arguments if a mathematical system makes sense at all. Maybe the adding-up-to-normality-ness of my system will become clearer, after it becomes clear that you can always isolate the framework-of-trusted-arguments of a human having a moral argument.

---

<sup>13</sup>. [http://lesswrong.com/lw/t6/the\\_cartoon\\_guide\\_to\\_G%C3%B6dels\\_theorem/](http://lesswrong.com/lw/t6/the_cartoon_guide_to_G%C3%B6dels_theorem/)

## 37. No License To Be Human<sup>1</sup>

**Followup to:** You Provably Can't Trust Yourself<sup>2</sup>

Yesterday<sup>3</sup> I discussed the difference between:

- A system that believes—is moved by—any *specific* chain of deductions from the axioms of Peano Arithmetic. (PA, Type 1 calculator)
- A system that believes PA, plus *explicitly* asserts the *general* proposition that PA is sound. (PA+1, meta-1-calculator that calculates the output of Type 1 calculator)
- A system that believes PA, plus explicitly asserts *its own* soundness. (Self-PA, Type 2 calculator)

These systems are formally distinct. PA+1 can prove things that PA cannot. Self-PA is inconsistent, and can prove anything via Löb's Theorem<sup>4</sup>.

With these distinctions in mind, I hope my intent will be clearer, when I say that although I am human and have a human-ish moral framework, I do not think that the fact of *acting in a human-ish* way licenses anything.

I am a self-renormalizing moral system, but I do not think there is any general license to be a self-renormalizing moral system.

And while we're on the subject, I am an epistemologically incoherent creature, trying to modify his ways of thinking in accordance with his current conclusions; but I do not think that reflective coherence implies correctness.

Let me take these issues in reverse order, starting with the general unlicensure of epistemological reflective coherence.

If five different people go out and investigate a city, and draw five different street maps, we should expect the maps to be (mostly roughly) consistent with each other. Accurate maps are necessarily consistent among each other and among themselves, there being only one reality. But if I sit in my living room with my blinds closed, I can draw up one street map from my imagination and then make four copies: these five maps will be consistent among them-

---

1. [http://lesswrong.com/lw/tg/no\\_license\\_to\\_be\\_human/](http://lesswrong.com/lw/tg/no_license_to_be_human/)

2. Page 1042, 'You Provably Can't Trust Yourself.'

3. Page 1042, 'You Provably Can't Trust Yourself.'

4. [http://lesswrong.com/lw/t6/the\\_cartoon\\_guide\\_to\\_lobs\\_theorem/](http://lesswrong.com/lw/t6/the_cartoon_guide_to_lobs_theorem/)

selves, but not accurate. Accuracy implies consistency but not the other way around.

In Where Recursive Justification Hits Bottom<sup>5</sup>, I talked about whether “I believe that induction will work on the next occasion, because it’s usually worked before” is legitimate reasoning, or “I trust Occam’s Razor because the simplest explanation for why Occam’s Razor often works is that we live in a highly ordered universe”. Though we actually *formalized* the idea of scientific induction, starting from an inductive *instinct*; we modified our intuitive understanding of Occam’s Razor (Maxwell’s Equations are in fact simpler than Thor, as an explanation for lightning) based on the simple idea that “the universe runs on equations, not heroic mythology”. So we did not automatically and unthinkingly confirm our assumptions, but rather, *used* our intuitions to *correct* them—seeking reflective coherence.

But I also remarked:

“And what about trusting reflective coherence in general? Wouldn’t most possible minds, randomly generated and allowed to settle into a state of reflective coherence, be incorrect? Ah, but *we* evolved by natural selection; we were not generated randomly.”

So you are not, *in general*, safe if you reflect on yourself and achieve internal coherence. The Anti-Inductors who compute that the probability of the coin coming up heads on the next occasion, decreases each time they see the coin come up heads, may defend their anti-induction by saying: “But it’s never worked before!”

The only reason why our human reflection works, is that we are good enough to make ourselves better—that we had a core instinct of induction, a core instinct of simplicity, that wasn’t sophisticated or exactly right, but worked well enough.

A mind that was completely wrong to start with, would have no seed of truth from which to heal itself. (It can’t forget everything and become a mind of pure emptiness that would mysteriously do induction correctly.)

So it’s not that reflective coherence is licensed *in general*, but that it’s a good idea *if* you start out with a core of truth or correctness or good priors. Ah, but who is deciding whether I possess good priors? I am! By reflecting on them! The inescapability of

---

<sup>5</sup>. Page 887, ‘Where Recursive Justification Hits Bottom’.

this strange loop is *why* a broken mind can't heal itself—because there is no jumping outside of *all* systems.

I can only plead that, in evolving to perform induction rather than anti-induction, in evolving a flawed but not absolutely wrong instinct for simplicity, I have been blessed with an epistemic gift.

I can only plead that self-renormalization works when *I* do it, even though it wouldn't work for Anti-Inductors. I can only plead that when *I* look over my flawed mind and see a core of useful reasoning, that I am really right, even though a completely broken mind might mistakenly perceive a core of useful truth.

Reflective coherence isn't licensed for all minds. It works for me, because I started out with an epistemic gift.

It doesn't matter if the Anti-Inductors look over themselves and decide that their anti-induction also constitutes an epistemic gift; they're wrong, *I'm* right.

And if that sounds philosophically indefensible, I beg you to step back from philosophy, and consider whether what I have just said is really truly *true*.

(Using your own concepts of induction and simplicity to do so, of course.)

Does this sound a little less indefensible, if I mention that PA trusts only proofs from the PA axioms, not proofs from every possible set of axioms? To the extent that I trust things like induction and Occam's Razor, then of course I don't trust anti-induction or anti-Occamian priors—*they wouldn't start working just because I adopted them*.

*What I trust* isn't a ghostly variable-framework from which I arbitrarily picked one possibility, so that picking any other would have worked as well so long as I renormalized it. *What I trust* is induction and Occam's Razor, which is why I use them to think about induction and Occam's Razor.

(Hopefully I have not just licensed myself to trust myself; only licensed being moved by both implicit and explicit appeals to induction and Occam's Razor. Hopefully this makes me PA+I, not Self-PA.)

So there is no *general, epistemological* license to be a self-renormalizing factual reasoning system.

The reason my system works is because it started out fairly inductive—not because of the naked meta-fact that it's trying to renormalize itself using *any* system; only induction counts. The license—no, the *actual usefulness*—comes from the inductive-ness, not from mere reflective-ness. Though I'm an inductor who says so!

And, sort-of similarly, but not exactly analogously:

There is no general *moral* license to be a self-renormalizing decision system. Self-consistency in your decision algorithms is not that-which-is-right.

The Pebblesorters<sup>6</sup> place the entire meaning of their lives in assembling correct heaps of pebbles and scattering incorrect ones; they don't know what makes a heap correct or incorrect, but they know it when they see it. It turns out that prime heaps are correct, but determining primality is not an easy problem for their brains. Like PA and unlike PA+1, the Pebblesorters are moved by particular and specific arguments tending to show that a heap is correct or incorrect (that is, prime or composite) but they have no explicit notion of “prime heaps are correct” or even “Pebblesorting People can tell which heaps are correct or incorrect”. They just know (some) correct heaps when they see them, and can try to figure out the others.

Let us suppose by way of supposition, that when the Pebblesorters are presented with the essence of their decision system—that is, the primality test—they recognize it with a great leap of relief and satisfaction. We can spin other scenarios—Peano Arithmetic, when presented with itself, does not prove itself correct. But let's suppose that the Pebblesorters recognize a wonderful method of systematically producing correct pebble heaps. Or maybe they don't endorse Adleman's test<sup>7</sup> as being the essence of correctness—any more than Peano Arithmetic proves that what PA proves is true—but they do recognize that Adleman's test is a wonderful way of producing correct heaps.

Then the Pebblesorters have a reflectively coherent decision system.

But this does not constitute a disagreement between them and humans about what is *right*, any more than humans, in scattering

6. Page 1008, ‘Sorting Pebbles Into Correct Heaps’.

7. [http://en.wikipedia.org/wiki/Primality\\_test](http://en.wikipedia.org/wiki/Primality_test)

a heap of 3 pebbles, are disagreeing with the Pebblesorters about which numbers are prime!

The Pebblesorters are moved by arguments like “Look at this row of 13 pebbles, and this row of 7 pebbles, arranged at right angles to each other; how can you see that, and still say that a heap of 91 pebbles is correct?”

Human beings are moved by arguments like “Hatred leads people to play purely negative-sum games, sacrificing themselves and hurting themselves to make others hurt still more” or “If there is not the threat of retaliation, carried out even when retaliation is profitless, there is no credible deterrent against those who can hurt us greatly for a small benefit to themselves”.

This is not a minor difference of flavors. When you reflect on the kind of arguments involved here, you are likely to conclude that the Pebblesorters *really are* talking about primality, whereas the humans *really are* arguing about what’s right. And I agree with this, since I am not a moral relativist. I don’t think that morality being moral implies any ontologically basic physical rightness attribute of objects; and conversely, I don’t think the lack of such a basic attribute is a reason to panic.

I may have contributed to the confusion here by labeling the Pebblesorters’ decisions “p-right”. But what they are talking about is not a different brand of “right”. What they’re talking about is prime numbers. There is no general rule that reflectively coherent decision systems are *right*; the Pebblesorters, in *merely* happening to implement a reflectively coherent decision system, are not yet talking about *morality*!

It’s been suggested that I should have spoken of “p-right” and “h-right”, not “p-right” and “right”.

But of course I made a very deliberate decision not to speak of “h-right”. That sounds like there is a general license to be human.

It sounds like being human is the essence of rightness. It sounds like the justification framework is “this is what humans do” and not “this is what saves lives, makes people happy, gives us control over our own lives, involves us with others and prevents us from collapsing into total self-absorption, keeps life complex and non-repeating and aesthetic and interesting, dot dot dot etcetera etcetera”.

It's possible that the above value list, or your equivalent value list, may not sound like a compelling notion unto you. Perhaps you are only moved to perform *particular* acts that make people happy—not caring all that much yet about this general, explicit, verbal notion of “making people happy is a value”. Listing out your values may not seem very valuable to you. (And I’m not even arguing with that judgment, in terms of everyday life; but a Friendly AI researcher has to know the metaethical score, and you may have to judge whether funding a Friendly AI project will make your children happy.) Which is just to say that you’re behaving like PA, not PA+I.

And as for that value framework being valuable because it’s human—why, it’s just the other way around: humans have received a *moral gift*, which Pebblesorters lack, in that we started out interested in things like happiness instead of just prime pebble heaps.

Now this is not actually a case of someone reaching in from outside with a gift-wrapped box; any more than the “moral miracle<sup>8</sup>” of blood-soaked natural selection producing Gandhi, is a real miracle.

It is only when you look out from *within* the perspective of morality, that it seems like a great wonder that natural selection could produce true friendship. And it is only when you look out from within the perspective of morality, that it seems like a great blessing that there are humans around to colonize the galaxies and do something interesting with them. From a purely causal perspective, nothing unlawful has happened.

But from a moral perspective, the wonder is that there are these human brains around that happen to *want* to help each other—a great wonder indeed, since human brains don’t *define* rightness, any more than natural selection *defines* rightness.

And that’s why I object to the term “h-right”. *I am not trying to do what’s human. I am not even trying to do what is reflectively coherent for me. I am trying to do what’s right.*

It may be that humans argue about what’s right, and Pebblesorters do what’s prime. But this doesn’t change what’s right, and it doesn’t make what’s right vary from planet to planet, and it doesn’t mean that the things we do are right in mere virtue of our deciding

---

8. Page 927, ‘The Gift We Give To Tomorrow’.

on them—any more than Pebblesorters *make* a heap prime or not prime by deciding that it's "correct".

The Pebblesorters aren't trying to do what's p-prime any more than humans are trying to do what's h-prime. The Pebblesorters are trying to do what's prime. And the humans are arguing about, and occasionally even really trying to do, what's right.

The Pebblesorters are not trying to create heaps of the sort that a Pebblesorter would create (note circularity). The Pebblesorters don't think that Pebblesorting thoughts have a special and supernatural influence on whether heaps are prime. The Pebblesorters aren't *trying* to do anything explicitly related to Pebblesorters—just like PA isn't trying to prove anything explicitly related to proof. PA just talks about numbers; it took a special and additional effort to encode any notions of proof in PA, to make PA talk about itself.

PA doesn't ask explicitly whether a theorem is provable in PA, before accepting it—indeed PA wouldn't care if it did prove that an encoded theorem was provable in PA. Pebblesorters don't care what's p-prime, just what's prime. And I don't give a damn about this "h-rightness" stuff: there's no license to be human, and it doesn't justify anything.

## 38. Invisible Frameworks<sup>1</sup>

**Followup to:** Passing the Recursive Buck<sup>2</sup>, No License To Be Human<sup>3</sup>

Roko has mentioned his “Universal Instrumental Values” several times in his comments. Roughly, Roko proposes that we ought to adopt as terminal values<sup>4</sup> those things that a supermajority of agents would do instrumentally<sup>5</sup>. On Roko’s blog he writes:

I’m suggesting that UIV provides the cornerstone for a rather new approach to goal system design. Instead of having a fixed utility function/supergoal, you periodically promote certain instrumental values to terminal values i.e. you promote the UIVs.

Roko thinks his morality is more *objective* than mine:

It also worries me quite a lot that eliezer’s post is entirely symmetric under the action of replacing his chosen notions with the pebble-sorter’s notions. This property qualifies as “moral relativism” in my book, though there is no point in arguing about the meanings of words.

My posts on universal instrumental values are not symmetric under replacing UIVs with some other set of goals that an agent might have. UIVs are the unique set of values X such that in order to achieve any other value Y, you first have to do X.

Well, and this proposal has a number of problems, as some of the commenters on Roko’s blog point out.

For a start, Roko actually says “universal”, not “supermajority”, but there are no actual universal actions; no matter what the green button *does*, there are possible mind designs whose utility function just says “Don’t press the green button.” There is no button, in

1. [http://lesswrong.com/lw/ta/invisible\\_frameworks/](http://lesswrong.com/lw/ta/invisible_frameworks/)

2. [http://lesswrong.com/lw/rd/passing\\_the\\_recursive\\_buck/](http://lesswrong.com/lw/rd/passing_the_recursive_buck/)

3. Page 1047, ‘No License To Be Human’.

4. [http://lesswrong.com/lw/l4/terminal\\_values\\_and\\_instrumental\\_values/](http://lesswrong.com/lw/l4/terminal_values_and_instrumental_values/)

5. [http://lesswrong.com/lw/l4/terminal\\_values\\_and\\_instrumental\\_values/](http://lesswrong.com/lw/l4/terminal_values_and_instrumental_values/)

other words, that all possible minds will press. Still, if you defined some prior weighting over the space of possible minds, you could probably find buttons that a supermajority would press, like the “Give me free energy” button.

But to do *nothing* except press such buttons, consists of constantly losing your purposes<sup>6</sup>. You find that driving the car is useful for getting and eating chocolate, or for attending dinner parties, or even for buying and manufacturing more cars. In fact, you realize that *every* intelligent agent will find it useful to travel places. So you start driving the car around without any destination. Roko hasn’t noticed this because, by anthropomorphic optimism<sup>7</sup>, he mysteriously only thinks of humanly appealing “UIVs” to propose, like “creativity”.

Let me guess, Roko, you don’t think that “drive a car!” is a “valid” UIV for some reason? But you did not apply some fixed procedure you had previously written down, to decide whether “drive a car” was a valid UIV or not. Rather you started out feeling a moment of initial discomfort, and then looked for reasons to disapprove. I wonder why the same discomfort didn’t occur to you when you considered “creativity”.

But let us leave aside the universality, appeal, or well-specifiedness of Roko’s metaethics.

Let us consider only Roko’s claim that his morality is more *objective* than, say, mine, or this marvelous list by William Frankena that Roko quotes SEP<sup>8</sup> quoting:

Life, consciousness, and activity; health and strength;  
pleasures and satisfactions of all or certain kinds;  
happiness, beatitude, contentment, etc.; truth;  
knowledge and true opinions of various kinds,  
understanding, wisdom; beauty, harmony, proportion in  
objects contemplated; aesthetic experience; morally good  
dispositions or virtues; mutual affection, love, friendship,  
cooperation; just distribution of goods and evils;  
harmony and proportion in one’s own life; power and  
experiences of achievement; self-expression; freedom;

---

6. [http://lesswrong.com/lw/le/lost\\_purposes/](http://lesswrong.com/lw/le/lost_purposes/)

7. [http://lesswrong.com/lw/st/anthropomorphic\\_optimism/](http://lesswrong.com/lw/st/anthropomorphic_optimism/)

8. <http://plato.stanford.edu/entries/value-intrinsic-extrinsic/#WbaHasIntVal>

peace, security; adventure and novelty; and good reputation, honor, esteem, etc.

So! Roko prefers his Universal Instrumental Values to this, because:

It also worries me quite a lot that eliezer's post is entirely symmetric under the action of replacing his chosen notions with the pebble-sorter's notions. This property qualifies as "moral relativism" in my book, though there is no point in arguing about the meanings of words.

My posts on universal instrumental values are not symmetric under replacing UIVs with some other set of goals that an agent might have. UIVs are the unique set of values X such that in order to achieve any other value Y, you first have to do X.

It would seem, then, that Roko attaches tremendous *importance* to claims to asymmetry and uniqueness; and tremendous disaffection to symmetry and relativism.

Which is to say that, when it comes to metamoral arguments, Roko is greatly moved to adopt morals by the statement "this goal is universal", while greatly moved to reject morals by the statement "this goal is relative".

In fact, so strong is this tendency of Roko's, that the metamoral argument "Many agents will do X!" is sufficient for Roko to adopt X as a terminal value. Indeed, Roko thinks that we ought to get *all* our terminal values this way.

Is this objective?

Yes and no.

When you evaluate the question "How many agents do X?", the answer does not depend on which agent evaluates it. It does depend on quantities like your weighting over all possible agents, and on the particular way you slice up possible events into categories<sup>9</sup> like "X". But let us be charitable: if you adopt a fixed weighting over agents and a fixed set of category boundaries, the question

---

9. Page 213, 'Disguised Queries'.

“How many agents do X?” has a unique answer. In this sense, Roko’s meta-utility function is objective.

But of course Roko’s meta-utility function is not “objective” in the sense of universal compellingness. It is only Roko who finds the argument “Most agents do X instrumentally” a compelling reason to promote X to a terminal value. I don’t find it compelling; it looks to me like losing purpose and double-counting expected utilities. The vast majority of possible agents, in fact, will not find it a compelling argument! A paperclip maximizer perceives no utility-function-changing, metamoral valence in the proposition “Most agents will find it useful to travel from one place to another.”

Now this seems like an extremely obvious criticism of Roko’s theory. Why wouldn’t Roko have thought of it?

Because when Roko feels like he’s being *objective*, he’s *using* his meta-morality as a fixed given—evaluating the question “How many agents do X?” in different places and times, but not asking any different questions. The answer to his meta-moral question has occurred to him as a variable to be investigated; the meta-moral question itself is off the table.

But—of course—when a Pebblesorter<sup>10</sup> regards “13 and 7!” as a powerful metamoral argument that “heaps of 91 pebbles” should not be a positive value in their utility function, they are asking a question whose answer is the same in all times and all places. They are asking whether 91 is prime or composite. A Pebblesorter, perhaps, would feel the same powerful surge of objectivity that Roko feels when Roko asks the question “How many agents have this instrumental value?” But in this case it readily occurs to Roko to ask “Why care if the heap is prime or not?” As it does not occur to Roko to ask, “Why care if this instrumental goal is universal or not?” Why... isn’t it just *obvious* that it matters whether an instrumental goal is universal?

The Pebblesorter’s framework is readily visible to Roko, since it differs from his own. But when Roko asks his own question—“Is this goal universally instrumental?”—he sees only the answer, and not the question; he sees only the output as a potential variable, not the framework.

---

<sup>10</sup>. Page 1008, ‘Sorting Pebbles Into Correct Heaps’.

Like PA<sup>11</sup>, that only sees the compellingness of particular proofs that use the Peano Axioms, and does not consider the quoted Peano Axioms as subject matter. It is only PA+I that sees the framework of PA.

But there is always a framework, every time you are moved to change your morals—the question is whether it will be invisible to you or not. That framework is always implemented in some particular brain, so that the same argument would fail to compel a differently constructed brain—though this does not imply that the framework makes any mention of brains at all.

And this difficulty of the invisible framework is at work, every time someone says, “But of course the correct morality is just *the one that helps you survive / the one that helps you be happy*”—implicit there is a supposed framework of meta-moral arguments that move you. But maybe I don’t think that being happy is the one and only argument that matters.

Roko is adopting a special and unusual metamoral framework in regarding “Most agents do X!” as a compelling reason to change one’s utility function. Why might Roko find this appealing? Humans, for very understandable reasons of evolutionary psychology, have a universalizing instinct<sup>12</sup>; we think that a valid argument should persuade anyone.

But what happens if we confess that such thinking can be valid? What happens if we confess that a meta-moral argument can (in its invisible framework) use the universalizing instinct? Then we have... just done something very human. We haven’t explicitly adopted the rule that all human instincts are good because they are human—but we did use one human instinct to think about morality. We didn’t explicitly think that’s what we were doing, any more than PA quotes itself in every proof; but we felt that a universally instrumental goal had this appealing quality of objective-ness about that, which is a perception of an intuition that evolved. This doesn’t mean that objective-ness is subjective. If you define objectiveness precisely then the question “What is objective?” will have a unique answer. But it does mean that we have just been compelled by an argument that will not compel every possible mind.

---

<sup>11</sup>. Page 1047, ‘No License To Be Human’.

<sup>12</sup>. Page 992, ‘Interpersonal Morality’.

If it's okay to be compelled by the appealing objectiveness of a moral, then why not also be compelled by...

...life, consciousness, and activity; health and strength; pleasures and satisfactions of all or certain kinds; happiness, beatitude, contentment, etc.; truth; knowledge and true opinions of various kinds, understanding, wisdom...

Such values, if precisely defined, can be just as objective as the question "How many agents do X?" in the sense that "How much health is in this region here?" will have a single unique answer. But it is humans who care about health, just as it is humans who care about universalizability.

The framework by which we care about health and happiness, as much evolved, and human, and part of the very substance of that which we name *right* whether it is human or not... as our tendency to find universalizable morals appealing.

And every sort of thing that a mind can do will have some framework behind it. Every sort of argument that can compel one mind, will fail to be an argument in the framework of another.

We are in the framework we name *right*; and every time we try to do what is *correct*, what we *should*, what we *must*, what we *ought*, that is the question we are asking.

Which question *should* we ask? What is the *correct* question?

Don't let your framework to *those* questions be invisible! Don't think you've answered them without asking any questions!

There is always the meta-meta-meta-question and it always has a framework.

I, for one, have decided to answer such questions the *right* way, as the alternative is to answer it the *wrong* way, like Roko is doing.

And the Pebblesorters do not disagree with any of this; they do what is objectively prime, not what is objectively right. And the Roko-AI does what is objectively often-instrumental, flying starships around with no destination; I don't disagree that travel is often-instrumental, I just say it is not right.

There is no right-ness that isn't in any framework—no feeling of rightness, no internal label that your brain produces, that can be detached from any method whatsoever of computing it—that just isn't what we're talking about when we ask "What should I do

now?" Because if anything labeled *should*, is *right*, then *that* is Self-PA.