Map and Territory

Eliezer Yudkowsky

2006 - 2009

Contents

The Simple Truth

> "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 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 benevolent 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 substrate. 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 *about-ness* 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 magic 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 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'."

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?

What Do We Mean By "Rationality"?

We mean:

1. **Epistemic rationality**: believing, and updating on evidence, so as to systematically improve the correspondence between your map and the territory. 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, 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("Bill plays jazz") $<$ P("Bill plays jazz" & "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:

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 in which the number of observers is uncertain. The exact rules of decision theory are called into question by e.g. Newcomblike problems 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 and here.)

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. Here again, trying to define "rationality" a particular way doesn't support an answer, merely presume it.

From 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.
>
> 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, not even if that word is "rationality". The point of attaching sequences of letters to particular concepts is to let two people *communicate* - to help transport thoughts from one mind to another. You cannot change reality, or prove the thought, by manipulating which meanings go with which words.

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

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

For example, if you say, "The rational belief is X, but the true belief is Y" then you are probably using the word "rational" in a way that means something other than what most of us have in mind. (E.g. some of us expect "rationality" to be *consistent under reflection* - "rationally" looking at the evidence, and "rationally" considering how your mind processes the evidence, shouldn't lead

to two different conclusions.) Similarly, if you find yourself saying "The rational thing to do is X, but the right thing to do is Y" then you are almost certainly using one of the words "rational" or "right" in a way that a huge chunk of readers won't agree with.

In this case - or in any other case where controversy threatens - you should substitute more specific language: "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, 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.

Why Truth? And...

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.) 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 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...

What's a Bias Again?

A *bias* is a certain kind of obstacle to our goal of obtaining truth - its character as an "obstacle" stems from this goal of truth - but there are many obstacles that are not "biases".

If we start right out by asking "What is bias?", it comes at the question in the wrong order. As the proverb goes, "There are forty kinds of lunacy but only one kind of common sense." The truth is a narrow target, a small region of configuration space to hit. "She loves me, she loves me not" may be a binary question, but E=MCˆ2 is a tiny dot in the space of all equations, like a winning lottery ticket in the space of all lottery tickets. Error is not an exceptional condition; it is success which is *a priori* so improbable that it requires an explanation.

We don't start out with a moral duty to "reduce bias", because biases are bad and evil and Just Not Done. This is the sort of thinking someone might end up with if they acquired a deontological duty of "rationality" by social osmosis, which leads to people trying to execute techniques without appreciating the

reason for them. (Which is bad and evil and Just Not Done, according to *Surely You're Joking, Mr. Feynman*, which I read as a kid.)

Rather, we want to get to the truth, for whatever reason, and we find various obstacles getting in the way of our goal. These obstacles are not wholly dissimilar to each other - for example, there are obstacles that have to do with not having enough computing power available, or information being expensive. It so happens that a large group of obstacles seem to have a certain character in common - to cluster in a region of obstacle-to-truth space - and this cluster has been labeled "biases".

What is a bias? Can we look at the empirical cluster and find a compact test for membership? Perhaps we will find that we can't really give any explanation better than pointing to a few extensional examples, and hoping the listener understands. If you are a scientist just beginning to investigate fire, it might be a lot wiser to point to a campfire and say "Fire is that orangey-bright hot stuff over there," rather than saying "I define fire as an alchemical transmutation of substances which releases phlogiston." As I said in The Simple Truth, you should not ignore something just because you can't define it. I can't quote the equations of General Relativity from memory, but nonetheless if I walk off a cliff, I'll fall. And we can say the same of biases - they won't hit any less hard if it turns out we can't define compactly what a "bias" is. So we might point to conjunction fallacies, to overconfidence, to the availability and representativeness heuristics, to base rate neglect, and say: "Stuff like that."

With all that said, we seem to label as "biases" those obstacles to truth which are produced, not by the cost of information, nor by limited computing power, but by the shape of our own mental machinery. For example, the machinery is evolutionarily optimized to purposes that actively oppose epistemic accuracy; for example, the machinery to win arguments in adaptive political contexts. Or the selection pressure ran skew to epistemic accuracy; for example, believing what others believe, to get along socially. Or, in the classic heuristic-and-bias, the machinery operates by an identifiable algorithm that does some useful work but also produces systematic errors: the availability heuristic is not itself a bias, but it gives rise to identifiable, compactly describable biases. Our brains are doing something wrong, and after a lot of experimentation and/or heavy thinking, someone identifies the problem in a fashion that System 2 can comprehend; then we call it a "bias". Even if we can do no better for knowing, it is still a failure that arises, in an identifiable fashion, from a particular kind of cognitive machinery - not from having too little machinery, but from the shape of the machinery itself.

"Biases" are distinguished from errors that arise from cognitive content, such as adopted beliefs, or adopted moral duties. These we call "mistakes", rather than "biases", and they are much easier to correct, once we've noticed them for ourselves. (Though the source of the mistake, or the source of the source of the mistake, may ultimately be some bias.)

"Biases" are distinguished from errors that arise from damage to an individual human brain, or from absorbed cultural mores; biases arise from machinery that is humanly universal.

Plato wasn't "biased" because he was ignorant of General Relativity - he had no way to gather that information, his ignorance did not arise from the shape of his mental machinery. But if Plato believed that philosophers would make better kings because he himself was a philosopher - and this belief, in turn, arose because of a universal adaptive political instinct for self-promotion, and not because Plato's daddy told him that everyone has a moral duty to promote their own profession to governorship, or because Plato sniffed too much glue as a kid - then that was a bias, whether Plato was ever warned of it or not.

Biases may not be cheap to correct. They may not even be correctable. But where we look upon our own mental machinery and see a causal account of an identifiable class of errors; and when the problem seems to come from the evolved shape of the machinery, rather from there being too little machinery, or bad specific content; then we call that a bias.

Personally, I see our quest in terms of acquiring personal skills of rationality, in improving truthfinding technique. The challenge is to attain the positive goal of truth, not to avoid the negative goal of failure. Failurespace is wide, infinite errors in infinite variety. It is difficult to describe so huge a space: "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." Success-space is narrower, and therefore more can be said about it.

While I am not averse (as you can see) to discussing definitions, we should remember that is not our primary goal. We are here to pursue the great human quest for truth: for we have desperate need of the knowledge, and besides, we're curious. To this end let us strive to overcome whatever obstacles lie in our way, whether we call them "biases" or not.

What is Evidence?

> "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. Today I'm going to talk about "evidence". (I also intend to discuss beliefs-of-fact, not emotions or morality, as distinguished here.)

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 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.

How Much Evidence Does It Take?

Previously, 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 in order to win? Suppose there are seventy balls, drawn without replacement, and six numbers to match for the win. Then there are 131,115,985 possible winning combinations, hence a randomly selected ticket would have a 1/131,115,985 probability of winning

(0.0000007%). To win the lottery, you would need evidence *selective* enough to visibly favor one combination over 131,115,984 alternatives.

Suppose there are some tests you can perform which discriminate, probabilistically, between winning and losing lottery numbers. For example, you can punch a combination into a little black box that always beeps if the combination is the winner, and has only a 1/4 (25%) chance of beeping if the combination is wrong. In Bayesian 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 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.

How To Convince Me That 2 + 2 = 3

In "What is Evidence?", I wrote:

> 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:

> 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.

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.

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 - either the belief got there via a mind-reality entangling process, or not. If not, the belief can't be correct except by coincidence. For beliefs with the slightest shred of internal complexity (requiring a computer program of more than 10 bits to simulate), the space of possibilities is large enough that coincidence vanishes.

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

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

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

Occam's Razor

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, 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 complexity of anger, and indeed the complexity of intelligence, was glossed over by the humans who hypothesized Thor the thunder-agent.

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

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

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

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

In the better (IMHO) versions of Solomonoff Induction, the computer program does not produce a deterministic prediction, but assigns probabilities to strings. For example, we could write a program to explain a fair coin by writing

a program that assigns equal probabilities to all 2^N strings of length N. This is Solomonoff Induction's approach to *fitting* the observed data. The higher the probability a program assigns to the observed data, the better that program *fits* the data. And probabilities must sum to 1, so for a program to better "fit" one possibility, it must steal probability mass from some other possibility which will then "fit" much more poorly. There is no superfair coin that assigns 100% probability to heads and 100% probability to tails.

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

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

The Lens That Sees Its Flaws

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, 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 at this, for it is indeed marvelous.

Mice see, but they don't know they have visual cortexes, so they can't correct for optical illusions. A mouse lives in a mental world that includes cats, holes, cheese and mousetraps - but not mouse brains. Their camera does not take pictures of its own lens. But we, as humans, can look at a seemingly bizarre image, 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 separate magisterium, far away from real life and the understanding of ordinary mortals. Science is not something that only applies to the inside of laboratories. 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 - this thought process will not systematically correlate 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 and measure the charge of entanglement quantitatively. Your optimistic nature cannot have *that* large an effect on the world; it cannot, of itself, decrease the probability of nuclear war by 20%, or however much your optimistic nature shifted your beliefs. Shifting your beliefs by a large amount, due to an event that only carries a very tiny charge of entanglement, will still mess up your mapping.)

To ask which beliefs make you happy, is to turn inward, not outward - it tells you something about yourself, but it is not evidence entangled with the environment. I have nothing anything against happiness, but it should follow from 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.