Mysterious Answers to Mysterious Questions

Eliezer Yudkowsky

July 28 - September 11, 2007

Contents

Making Beliefs Pay Rent (in Anticipated Experiences)

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 verbalish, propositional. It probably does not exaggerate much to describe these two beliefs as sentences made out of words. But these two beliefs have an inferential *consequence* that is a direct sensory anticipation - if the clock's second hand is on the 12 numeral when you drop the ball, you anticipate seeing it on the 1 numeral when you hear the crash five seconds later. To anticipate sensory experiences as precisely as possible, we must process beliefs that are not anticipations of sensory experience.

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

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

can call it that, doesn't connect to sensory experience at all. But you had better remember the propositional assertion that "Wulky Wilkinsen" has the "post-utopian" attribute, so you can regurgitate it on the upcoming quiz. Likewise if "post-utopians" show "colonial alienation"; if the quiz asks whether Wulky Wilkinsen shows colonial alienation, you'd better answer yes. The beliefs are connected to each other, though still not connected to any anticipated experience.

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

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

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

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

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

Belief in Belief

Carl Sagan once told a parable 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. The claimant clearly does not *anticipate* seeing anything unusual upon opening the garage door; otherwise he wouldn't make advance excuses. It may also be that the claimant's pool of propositional beliefs contains *There is a dragon in my garage.* It may seem, to a rationalist, that these two beliefs should collide and conflict even though they are of different types. Yet it is a physical fact that you can write "The sky is green!" next to a picture of a blue sky without the paper bursting into flames.

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

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

Depending on how your childhood went, you may remember a time period when you first began to doubt Santa Claus's existence, but you still believed that you were *supposed* to believe in Santa Claus, so you tried to deny the doubts. As Daniel Dennett observes, where it is difficult to believe a thing, it is often much easier to believe that you *ought* to believe it. What does it mean to believe that the Ultimate Cosmic Sky 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 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.

Bayesian Judo

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

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 requires." I stretched out my hand, and he shook it, and then he wandered away.

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

"Thank you very much," I said.

Professing and Cheering

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" 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 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" 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.) It wasn't just a cheer, like marching, but an outrageous cheer, like marching naked - believing that she couldn't be arrested or criticized, because she was doing it for her pride parade.

That's why it mattered to her that what she was saying was beyond ridiculous.

If she'd tried to make it sound more plausible, it would have been like putting on clothes.

Belief as Attire

I have so far distinguished between belief as anticipation-controller, belief in belief, professing and cheering. 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 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, 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. 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, or religious professing, 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 fortheir favorite sports team. (This is the foundation on which rests the swindle of "Republicans vs. Democrats" and analogous false dilemmas 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.

Focus Your Uncertainty

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 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. 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 exactly 42 minutes on excuse #3? Why not 41 minutes, or 43? Admit it - you're not being objective! You're playing subjective favorites!*

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

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

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

It even seems like there's a relation between how much you anticipate each of the three outcomes, and how much time you want to spend preparing each excuse. Of course the relation can't actually be quantified. You have 100 minutes to prepare your speech, but there isn't 100 of anything to divide up in this anticipation business. (Although you do work out that, *if* some particular outcome occurs, then your utility function is logarithmic in time spent preparing the excuse.)

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

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

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

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

The Virtue of Narrowness

> 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.

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

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

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

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

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

It is perfectly all right for modern evolutionary biologists to explain *just* the patterns of living creatures, and not the "evolution" of stars or the "evolution" of technology. Alas, some unfortunate souls use the same word "evolution" to cover the naturally selected patterns of replicating life, *and* the strictly accidental structure of stars, *and* the intelligently configured structure of technology. And as we all know, if people use the same word, it must all be the same thing. You should automatically generalize anything you think you know about

biological evolution to technology. Anyone who tells you otherwise must be a mere pointless pedant. It couldn't possibly be that your abysmal ignorance of modern evolutionary theory is so total that you can't tell the difference between a carburetor and a radiator. That's unthinkable. No, the *other* guy - you know, the one who's studied the math - is just too dumb to see the connections.

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

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

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

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

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

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

> "If anyone should throw back his head and learn something by staring at the varied patterns on a ceiling, apparently you would think that he was contemplating with his reason, when he was only staring with his eyes... I cannot but believe that no study makes the soul look on high except that which is concerned with real being and the unseen. Whether he gape and stare upwards, or shut his mouth and stare downwards, if it be things of the senses that he tries to learn something about, I declare he never could learn, for none of these things admit of knowledge: I say his soul is looking down, not up, even if he is floating on his back on land or on sea!"

Many today make a similar mistake, and think that narrow concepts are as lowly and unlofty and unphilosophical as, say, going out and looking at things - an

endeavor only suited to the underclass. But rationalists - and also poets - need narrow words to express precise thoughts; they need categories which include only some things, and exclude others. There's nothing wrong with focusing your mind, narrowing your categories, excluding possibilities, and sharpening your propositions. Really, there isn't! If you make your words too broad, you end up with something that isn't true and doesn't even make good poetry.

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

Your Strength as a Rationalist

(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...

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 than a published journal article. Alas, belief is easier than disbelief; we believe instinctively, but disbelief requires a conscious effort.

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.

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."

Absence of Evidence is Evidence of Absence

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

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

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

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 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. 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.

Conservation of Expected Evidence

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" 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.

$$\mathbf{P(H) = P(H)}^*$$
$$P(H) = P(H,E) + P(H,{\sim}E)^*$$
$$\mathbf{P(H) = P(H|E)*P(E) + P(H|{\sim}E)*P({\sim}E)}$$

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

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

So if you claim 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 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. 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 oppositive expectation of counterevidence. If you try to weaken the counterevidence of a possible "abnormal" observation, you can only do it by weakening the support of a "normal" observation, to a precisely equal and opposite degree. It is a zero-sum game. No matter how you connive, no matter how you argue, no matter how you strategize, you can't possibly expect the resulting game plan to shift your beliefs (on average) in a particular direction.

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

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

Hindsight Bias

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

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

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

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

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

Fischhoff, B. 1982. For those condemned to study the past: Heuristics and biases in hindsight. In Kahneman et. al. 1982: 332â351.

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 â Ex Ante: Determining Liability in Hindsight. Law and Human Behavior, 19(1): 89–104.

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

Hindsight Devalues Science

This excerpt 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. (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? Take a moment to think before continuing...

In this demonstration (from Paul Lazarsfeld by way of Meyers), all of the findings above are the *opposite* of what was actually found. How many times did you think your model took a hit? How many times did you admit you would have been wrong? That's how good your model really was. The measure of your strength as a rationalist 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 thatdoesn't fit what you *really* would have expected.

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

Fake Explanations

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."

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? The words are easily enough professed - said in a loud, emphatic voice. But do the words actually control anticipation?

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.*

And as we all know by this point (I do hope), 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.

Guessing the Teachers Password

When I was young, I read popular physics books such as Richard Feynman's *QED: The Strange 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, 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. 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") is not a *hypothesis,* it is a pattern of vibrations traveling through the air, or ink on paper. It can *associate* to a hypothesis in someone's mind, but it is not, of itself, right or wrong. But in school, the teacher hands you a gold star for *saying* "made of waves", which must be the correct answer because the teacher heard a physicist emit the same sound-vibrations. Since verbal behavior (spoken or written) is what gets the gold star, students begin to think that verbal behavior has a truth-value. After all, either light is made of waves, or it isn't, right?

And this leads into an even worse habit. Suppose the teacher presents you with a confusing problem 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. 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 flowing-ness connects to experience; it controls anticipation.

If you aren't *using* the diffusion equation - putting in numbers and getting out results that control your anticipation of particular experiences - then the connection between map and territory is severed as though by a knife. What remains is not a belief, but a verbal behavior.
In the school system, it's all about verbal behavior, whether written on paper or spoken aloud. Verbal behavior gets you a gold star or a failing grade. Part of unlearning this bad habit is becoming consciously aware of the difference between an explanation and a password.

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

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

- "Heat conduction?"
- But that's only a phrase - what does it mean?

- The diffusion equation?
- But those are only symbols - how do I apply them?
- 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 that I am confused. Maybe the near side just *feels* cooler, because it's made of more insulative material and transfers less heat to my hand? I'll try measuring the temperature...
- Okay, that wasn't it. Can I try to verify whether the diffusion equation holds true of this metal plate, at all? Is heat *flowing* the way it usually does, or is something else going on?
- I could hold a match to the plate and try to measure how heat spreads over time...

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

(This post is part of the sequence Mysterious Answers to Mysterious Questions.)

Science as Attire

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

Probably an actual majority of the people who *believe in* evolution use the phrase "because of evolution" because they want to be part of the scientific in-crowd - belief as scientific attire, 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". 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 doesn't look to them like a lab coat; this isn't the football team they're cheering for.

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

Fake Causality

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; it could explain everything.

This was an earlier age of science. For a long time, no one realized there was a problem. Fake explanations 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]

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, it may feel like the cause constrains the effect, when it was merely* *fitted to the effect.

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

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

An analogous principle operates in rigorous probabilistic reasoning about causality. If you learn something about whether it's raining, from some source *other* than observing the sidewalk to be wet, this will send a forward-message from [rain] to [sidewalk wet] and raise our expectation of the sidewalk being wet. If you observe the sidewalk to be wet, this sends a backward-message to our belief that it is raining, and this message propagates from [rain] to all neighboring nodes *except* the [sidewalk wet] node. We count each piece of evidence exactly once; no update message ever "bounces" back and forth. The exact algorithm may be found in Judea Pearl's classic "Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference".

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

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

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

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

Speaking of "hindsight bias" 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, 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.

Semantic Stopsigns

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, which gives people a motive to say it as often as possible - some people even say it for questions like "Why did this hurricane strike New Orleans?" Even so, you'd hope people would notice that on the *particular* puzzle of the First Cause, saying "God!" doesn't help. It doesn't make the paradox seem any less paradoxical *even if true.* How could anyone *not* notice this?

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

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

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

(This post is part of the sequence Mysterious Answers to Mysterious Questions.)

Mysterious Answers to Mysterious Questions

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 - Whler's later synthesis of urea, a component of urine, was a major blow to the vitalistic theory because it showed that mere *chemistry* could duplicate a product of biology.

Calling "elan vital" an explanation, even a fake explanation like phlogiston, is probably giving it too much credit. It functioned primarily as a curiosity-stopper. 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 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 in their ignorance, made biology into a *sacred* mystery, and thereby became loath to relinquish their ignorance 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 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 - 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 rather than an anticipation-controller.
- 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 causethis or that; but the reason why the mysterious force behaves thus is wrapped in a blank unity.
- Third, those who proffer the explanation cherish their ignorance; 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.

The Futility of Emergence

The failures of phlogiston and vitalism are historical hindsight. 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: "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.

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. 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.

Another fun exercise is to replace the word "emergent" with the old word, the explanation 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?

"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 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 of settled science, they get up to the same shenanigans as their ancestors, dressed up in the literary genre of "science" but still the same species psychology.

Say Not "Complexity"

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?"

"Yes," said Marcello.

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

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

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

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

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

In an eyeblink it happens: putting a non-controlling causal node behind something mysterious, a causal node that feels like an explanation 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, since the ancient times.

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

I suspect that in academia there is a huge pressure to sweep problems under the rug so that you can present a paper with the appearance of completeness. You'll get more kudos for a seemingly complete model that includes some "emergent phenomena", 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 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, since skip-over happens by default 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 create an illusion of understanding. Wiser to say "magic", and leave yourself a placeholder, a reminder of work you will have to do later.

Positive Bias: Look Into the Dark

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.

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 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* - if you are equally good at explaining any outcome, you have zero knowledge. So to spot an explanation that isn't helpful, it's not enough to think of what it does explain very well - you also have to search for results it *couldn't* explain, and this is the true strength of the theory.

So I said all this, and then yesterday, I challenged the usefulness of "emergence" as a concept. 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?

My Wild and Reckless Youth

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. 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. 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 unto me, and possessed the same quality of wondrous impenetrability that it had at the start.

Make no mistake, 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.

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 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 it quantum gravity. But I knew not the Way of Bayes, and did not see the level 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. You need *one whole hell of a lot* of rationality before it does anything but lead you into new and interesting mistakes.* *

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

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

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

Failing to Learn from History

Once upon a time, in my wild and reckless youth, when I knew not the Way of Bayes, I gave a Mysterious Answer 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, a tendency toward mysterianism, which led them to come up with mysterious explanations for non-mysterious subjects. But surely, if a phenomenon really *was* very weird, a weird explanation might be in order?

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

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

Making History Available

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 - 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.

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. 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.

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

When I finally realized whose shoes I was standing in, 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 - *all* the memories, not just the memories which, by mere coincidence, belonged to myself and my own era.

The Earth became older, of a sudden.

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

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

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

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

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

Explain/Worship/Ignore?

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 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.

"Science" as Curiosity-Stopper

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

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

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

someone *else* knows, it devalues the knowledge in your eyes. You become less curious.

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

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

Besides, consider the consequences if you permit "Someone else knows the answer" to function as a curiosity-stopper. One 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, "SOMEONE ELSE ALREADY KNOWS WHY THIS ELEPHANT IS HERE."

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

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

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

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

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

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.

Applause Lights

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:

> 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 (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 propositional* *statement, as the equivalent of the "Applause" light that tells a studio audience when to clap.

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

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

If you reverse this statement, you get:

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

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

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

> I am here to propose to you today that we need to balance the risks and opportunities of advanced Artificial Intelligence. We should avoid the risks and, insofar as it is possible, realize the opportunities. We should not needlessly confront entirely unnecessary dangers. To achieve these goals, we must plan wisely and rationally. We should not act in fear and panic, or give in to technophobia; but neither should we act in blind enthusiasm. We should respect the interests of all parties with a stake in the Singularity. We must try to ensure that the benefits of advanced technologies accrue to as many individuals as possible, rather than being restricted to a few. We must try to avoid, as much as possible, violent conflicts using these technologies; and we must prevent massive destructive capability from falling into the hands of individuals. We should think through these issues before, not after, it is too late to do anything about them...

Chaotic Inversion

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 (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, right out there in my ordinary everyday life instead of high abstraction.

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

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

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

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".