Reductionism

Eliezer Yudkowsky

March 8 - April 5, 2008

Contents

Dissolving the Question **1**

Wrong Questions **4**

Righting a Wrong Question **6**

Mind Projection Fallacy **8**

Probability is in the Mind **9**

The Quotation is not the Referent **13**

Qualitatively Confused **15**

Reductionism **17**

Explaining vs Explaining Away **20**

Fake Reductionism **23**

Savanna Poets **25**

Joy in the Merely Real **28**

Joy in Discovery **29**

Bind Yourself to Reality **32**

If You Demand Magic, Magic Won't Help **34**

Mundane Magic **36**

The Beauty of Settled Science **39**

Amazing Breakthrough Day 1: April 1st **40**

Is Humanism a Religion-Substitute? **42**

Scarcity **44**

To Spread Science, Keep It Secret **46**

Initiation Ceremony **49**

Awww, a Zebra **51**

Hand vs Fingers **51**

Angry Atoms **54**

Heat vs Motion **57**

Brain Breakthrough! It's Made of Neurons! **60**

Reductive Reference **61**

Zombies! Zombies? **65**

Zombie Responses **80**

The Generalized Anti-Zombie Principle **85**

Gazp vs Glut **91**

Belief in the Implied Invisible **97**

Zombies: The Movie **101**

Excluding the Supernatural **105**

Psychic Powers **109**

Dissolving the Question

"If a tree falls in the forest, but no one hears it, does it make a sound?"

I didn't *answer* that question. I didn't pick a position, "Yes!" or "No!", and defend it. Instead I went off and deconstructed the human algorithm for processing words, even going so far as to sketch an illustration of a neural network. At the end, I hope, there was no question left - not even the feeling of a question.

Many philosophers - particularly amateur philosophers, and ancient philosophers - share a dangerous instinct: If you give them a question, they try to answer it.

Like, say, "Do we have free will?"

The dangerous instinct of philosophy is to marshal the arguments in favor, and marshal the arguments against, and weigh them up, and publish them in a prestigious journal of philosophy, and so finally conclude: "Yes, we must have free will," or "No, we cannot possibly have free will."

Some philosophers are wise enough to recall the warning that most philosophical disputes are really disputes over the meaning of a word, or confusions generated by using different meanings for the same word in different places. So they try to define very precisely what they mean by "free will", and then ask again, "Do we have free will? Yes or no?"

A philosopher wiser yet, may suspect that the confusion about "free will" shows the notion itself is flawed. So they pursue the Traditional Rationalist course: They argue that "free will" is inherently self-contradictory, or meaningless because it has no testable consequences. And then they publish these devastating observations in a prestigious philosophy journal.

But *proving that* you are confused may not make you feel any *less* confused. Proving that a question is meaningless may not help you any more than answering it.

The philosopher's instinct is to find the most defensible position, publish it, and move on. But the "naive" view, the instinctive view, is a fact about human psychology. You can prove that free will is impossible until the Sun goes cold, but this leaves an unexplained fact of cognitive science: If free will doesn't exist, what goes on inside the head of a human being who thinks it does? This is not a rhetorical question!

It is a fact about human psychology that people think they have free will. Finding a more defensible *philosophical position* doesn't change, or explain, that *psychological fact.* Philosophy may lead you to *reject* the concept, but rejecting a concept is not the same as understanding the cognitive algorithms behind it.

You could look at the Standard Dispute over "If a tree falls in the forest, and no one hears it, does it make a sound?", and you could do the Traditional Rationalist thing: Observe that the two don't disagree on any point of anticipated experience, and triumphantly declare the argument pointless. That happens to be correct in this particular case; but, as *a question of cognitive science,* why did the arguers make that mistake in the first place?

The key idea of the heuristics and biases program is that the *mistakes* we make, often reveal far more about our underlying cognitive algorithms than our correct answers. So (I asked myself, once upon a time) what kind of mind design corresponds to the mistake of arguing about trees falling in deserted forests?

The cognitive algorithms we use, *are* the way the world feels. And these cognitive algorithms may not have a one-to-one correspondence with reality - not even macroscopic reality, to say nothing of the true quarks. There can be things in the mind that cut skew to the world.

For example, there can be a dangling unit in the center of a neural network, which does not correspond to any real thing, or any real property of any real thing, existent anywhere in the real world. This dangling unit is often useful as a shortcut in computation, which is why we have them. (Metaphorically speaking. Human neurobiology is surely far more complex.)

This dangling unit *feels like* an unresolved question, even after every answerable query is answered. No matter how much anyone proves to you that no difference of anticipated experience depends on the question, you're left wondering: "But does the falling tree *really* make a sound, or not?"

But once you understand *in detail* how your brain generates the *feeling* of the question - once you realize that your feeling of an unanswered question, corresponds to an illusory central unit wanting to know whether it should fire, even after all the edge units are clamped at known values - or better yet, you understand the technical workings of Naive Bayes - *then* you're done. Then there's no lingering feeling of confusion, no vague sense of dissatisfaction.

If there is *any* lingering feeling of a remaining unanswered question, or of having been fast-talked into something, then this is a sign that you have not dissolved the question. A vague dissatisfaction should be as much warning as a shout. *Really* dissolving the question doesn't leave anything behind.

A triumphant thundering refutation of free will, an absolutely unarguable proof that free will cannot exist, feels very *satisfying* - a grand cheer for the home team. And so you may not notice that - as a point of cognitive science - you do not have a full and satisfactory descriptive explanation of how each intuitive sensation arises, point by point.

You may not even want to admit your ignorance, of this point of cognitive science, because that would feel like a score against Your Team. In the midst of smashing all foolish beliefs of free will, it would seem like a concession to the opposing side to concede that you've left anything unexplained.

And so, perhaps, you'll come up with a just-so evolutionary-psychological argument that hunter-gatherers who believed in free will, were more likely to take a positive outlook on life, and so outreproduce other hunter-gatherers - to give one example of a completely bogus explanation. If you say this, you are *arguing that* the brain generates an illusion of free will - but you are not *explaining how*. You are trying to dismiss the opposition by deconstructing its motives - but in the story you tell, the illusion of free will is a brute fact. You have not taken the illusion apart to see the wheels and gears.

Imagine that in the Standard Dispute about a tree falling in a deserted forest, you first prove that no difference of anticipation exists, and then go on to hypothesize, "But perhaps people who said that arguments were meaningless were viewed as having conceded, and so lost social status, so now we have an instinct to argue about the meanings of words." That's *arguing that* or *explaining why* a confusion exists. Now look at the neural network structure in Feel the Meaning. That's *explaining how*, disassembling the confusion into smaller pieces which are not themselves confusing. See the difference?

Coming up with good hypotheses about cognitive algorithms (or even hypotheses that hold together for half a second) is a good deal harder than just refuting a philosophical confusion. Indeed, it is an entirely different art. Bear this in mind, and you should feel less embarrassed to say, "I know that what you say can't possibly be true, and I can prove it. But I cannot write out a flowchart which shows how your brain makes the mistake, so I'm not done yet, and will continue investigating."

I say all this, because it sometimes seems to me that at least 20% of the real-world effectiveness of a skilled rationalist comes from not stopping too early. If you keep asking questions, you'll get to your destination eventually. If you decide too early that you've found an answer, you won't.

The challenge, above all, is to notice when you are confused - even if it just feels like a little tiny bit of confusion - and even if there's someone standing across from you, *insisting* that humans have free will, and *smirking* at you, and the fact that you don't know *exactly* how the cognitive algorithms work, has *nothing to do* with the searing folly of their position...

But when you can lay out the cognitive algorithm in sufficient detail that you can walk through the thought process, step by step, and describe how each intuitive perception arises - decompose the confusion into smaller pieces not themselves confusing - *then* you're done.

So be warned that you may *believe* you're done, when all you have is a mere triumphant refutation of a mistake.

But when you're *really* done, you'll *know* you're done. Dissolving the question is an unmistakable feeling - once you experience it, and, having experienced it, resolve not to be fooled again. Those who dream do not know they dream, but when you wake you know you are awake.

Which is to say: When you're done, you'll know you're done, but unfortunately the reverse implication does not hold.

So here's your homework problem: What kind of cognitive algorithm, as felt from the inside, would generate the observed debate about "free will"?

Your assignment is not to argue about whether people have free will, or not.

Your assignment is not to argue that free will is compatible with determinism, or not.

Your assignment is not to argue that the question is ill-posed, or that the concept is self-contradictory, or that it has no testable consequences.

You are not asked to invent an evolutionary explanation of how people who believed in free will would have reproduced; nor an account of how the concept of free will seems suspiciously congruent with bias X. Such are mere attempts to *explain why* people believe in "free will", not *explain how.*

Your homework assignment is to write a stack trace of the internal algorithms of the human mind as they produce the intuitions that power the whole damn philosophical argument.

This is one of the first real challenges I tried as an aspiring rationalist, once upon a time. One of the easier conundrums, relatively speaking. May it serve you likewise.

Wrong Questions

Where the mind cuts against reality's grain, it generates *wrong questions* - questions that cannot possibly be answered *on their own terms,* but only dissolved by understanding the cognitive algorithm that generates the *perception* of a question.

One good cue that you're dealing with a "wrong question" is when you cannot even *imagine* any concrete, specific state of how-the-world-is that would answer the question. When it doesn't even seem *possible* to answer the question.

Take the Standard Definitional Dispute, for example, about the tree falling in a deserted forest. Is there any way-the-world-could-be - any state of affairs - that corresponds to the word "sound" *really meaning* only acoustic vibrations, or *really meaning* only auditory experiences?

("Why, yes," says the one, "it is the state of affairs where 'sound' means acoustic vibrations." So Taboo the word 'means', and 'represents', and all similar

synonyms, and describe again: How can the world be, what state of affairs, would make one side right, and the other side wrong?)

Or if that seems too easy, take free will: What concrete state of affairs, whether in deterministic physics, or in physics with a dice-rolling random component, could ever correspond to having free will?

And if *that* seems too easy, then ask "Why does anything exist at all?", and then tell me what a satisfactory answer to that question would even *look like.*

And no, I don't know the answer to that last one. But I *can* guess one thing, based on my previous experience with unanswerable questions. The answer will not consist of some grand triumphant First Cause. The question will go away as a result of some insight into how my mental algorithms run skew to reality, after which I will understand how the question itself was wrong from the beginning - how the question itself assumed the fallacy, contained the skew.

Mystery exists in the mind, not in reality. If I am ignorant about a phenomenon, that is a fact about my state of mind, not a fact about the phenomenon itself. All the more so, if it seems like no possible answer can exist: Confusion exists in the map, not in the territory. *Unanswerable* questions do not mark places where magic enters the universe. They mark places where your mind runs skew to reality.

Such questions *must* be dissolved. Bad things happen when you try to answer them. It inevitably generates the worst sort of Mysterious Answer to a Mysterious Question: The one where you come up with seemingly strong arguments for your Mysterious Answer, but the "answer" doesn't let you make any new predictions even in retrospect, and the phenomenon still possesses the same sacred inexplicability that it had at the start.

I could guess, for example, that the answer to the puzzle of the First Cause is that nothing *does* exist - that the whole concept of "existence" is bogus. But if you sincerely believed that, would you be any less confused? Me neither.

But the wonderful thing about *unanswerable* questions is that they are *always* solvable, at least in my experience. What went through Queen Elizabeth I's mind, first thing in the morning, as she woke up on her fortieth birthday? As I can easily *imagine*answers to this question, I can readily see that I may never be able to *actually* answer it, the true information having been lost in time.

On the other hand, "Why does anything exist at all?" seems *so* absolutely impossible that I can infer that I am just confused, one way or another, and the truth probably isn't all that complicated in an absolute sense, and once the confusion goes away I'll be able to see it.

This may seem counterintuitive if you've never solved an unanswerable question, but I assure you that it *is* how these things work.

Coming tomorrow: A simple trick for handling "wrong questions".

Righting a Wrong Question

When you are faced with an *unanswerable* question - a question to which it seems impossible to even *imagine* an answer - there is a simple trick which can turn the question solvable.

Compare:

- "Why do I have free will?"
- "Why do I think I have free will?"

The nice thing about the second question is that it is *guaranteed* to have a real answer, *whether or not* there is any such thing as free will. Asking "Why do I have free will?" or "Do I have free will?" sends you off thinking about tiny details of the laws of physics, so distant from the macroscopic level that you couldn't begin to see them with the naked eye. And you're asking "Why is X the case?" where X may not be *coherent,* let alone the case.

"Why do I *think* I have free will?", in contrast, is guaranteed answerable. You do, in fact, believe you have free will. This belief seems far more solid and graspable than the ephemerality of free will. And there is, *in fact,* some nice solid chain of cognitive cause and effect leading up to this belief.

If you've already outgrown free will, choose one of these substitutes:

- "Why does time move forward instead of backward?" versus "Why do I think time moves forward instead of backward?"
- "Why was I born as myself rather than someone else?" versus "Why do I think I was born as myself rather than someone else?"
- "Why am I conscious?" versus "Why do I think I'm conscious?"
- "Why does reality exist?" versus "Why do I think reality exists?"

The beauty of this method is that it works *whether or not* the question is confused. As I type this, I am wearing socks. I could ask "Why am I wearing socks?" or "Why do I believe I'm wearing socks?" Let's say I ask the second question. Tracing back the chain of causality, I find:

- I believe I'm wearing socks, because I can see socks on my feet.
- I see socks on my feet, because my retina is sending sock signals to my visual cortex.
- My retina is sending sock signals, because sock-shaped light is impinging on my retina.

- Sock-shaped light impinges on my retina, because it reflects from the socks I'm wearing.
- It reflects from the socks I'm wearing, because I'm wearing socks.
- I'm wearing socks because I put them on.
- I put socks on because I believed that otherwise my feet would get cold.
- &c.

Tracing back the chain of causality, step by step, I discover that my belief that I'm wearing socks is fully explained by the fact that I'm wearing socks. This is right and proper, as you cannot gain information about something without interacting with it.

On the other hand, if I see a mirage of a lake in a desert, the correct causal explanation of my vision does not involve the fact of any actual lake in the desert. In this case, my belief in the lake is not just *explained,* but *explained away.*

But *either way,* the belief itself is a real phenomenon taking place in the real universe - psychological events are events - and its causal history can be traced back.

"Why is there a lake in the middle of the desert?" may fail if there is no lake to be explained. But "Why do I *perceive* a lake in the middle of the desert?" always has a causal explanation, one way or the other.

Perhaps someone will see an opportunity to be clever, and say: "Okay. I believe in free will because I have free will. There, I'm done." Of course it's not that easy.

My perception of socks on my feet, is an event in the visual cortex. The workings of the visual cortex can be investigated by cognitive science, should they be confusing.

My retina receiving light is not a mystical sensing procedure, a magical sock detector that lights in the presence of socks for no explicable reason; there are mechanisms that can be understood in terms of biology. The photons entering the retina can be understood in terms of optics. The shoe's surface reflectance can be understood in terms of electromagnetism and chemistry. My feet getting cold can be understood in terms of thermodynamics.

So it's not as easy as saying, "I believe I have free will because I have it - there, I'm done!" You have to be able to break the causal chain into smaller steps, and explain the steps in terms of elements not themselves confusing.

The mechanical interaction of my retina with my socks is quite clear, and can be described in terms of non-confusing components like photons and electrons. Where's the free-will-sensor in your brain, and how does it detect the

presence or absence of free will? How does the sensor interact with the sensed event, and what are the mechanical details of the interaction?

If your belief does derive from valid observation of a real phenomenon, we will eventually reach that fact, if we start tracing the causal chain backward from your belief.

If what you are really seeing is your own confusion, tracing back the chain of causality will find an algorithm that runs skew to reality.

Either way, the question is guaranteed to have an answer. You even have a nice, concrete place to begin tracing - your belief, sitting there solidly in your mind.

Cognitive science may not seem so lofty and glorious as metaphysics. But at least questions of cognitive science are *solvable.* Finding an answer may not be *easy,* but at least an answer *exists.*

Oh, and also: the idea that cognitive science is not so lofty and glorious as metaphysics is simply wrong. Some readers are beginning to notice this, I hope.

Mind Projection Fallacy

the dawn days of science fiction, alien invaders would occasionally kidnap a girl in a torn dress and carry her off for intended ravishing, as lovingly depicted on many ancient magazine covers. Oddly enough, the aliens never go after men in torn shirts.

Would a non-humanoid alien, with a different evolutionary history and evolutionary psychology, sexually desire a human female? It seems rather unlikely. To put it mildly.

People don't make mistakes like that by deliberately reasoning: "All possible minds are likely to be wired pretty much the same way, therefore a bug-eyed monster will find human females attractive." Probably the artist did not even think to ask whether an alien *perceives* human females as attractive. Instead, a human female in a torn dress *is sexy* - inherently so, as an intrinsic property.

They who went astray did not think about the alien's evolutionary history; they focused on the woman's torn dress. If the dress were not torn, the woman would be less sexy; the alien monster doesn't enter into it.

Apparently we instinctively represent Sexiness as a direct attribute of the Woman object, `Woman.sexiness`, like `Woman.height` or `Woman.weight`.

If your brain uses that data structure, or something metaphorically similar to it, then from the inside it feels like sexiness is an inherent property of the woman, not a property of the alien looking at the woman. Since the woman *is attractive,* the alien monster will be *attracted* to her - isn't that logical?

E. T. Jaynes used the term Mind Projection Fallacy to denote the error of projecting your own mind's properties into the external world. Jaynes, as a late grand master of the Bayesian Conspiracy, was most concerned with the mistreatment of *probabilities*as inherent properties of objects, rather than states of partial knowledge in some particular mind. More about this shortly.

But the Mind Projection Fallacy generalizes as an error. It is in the argument over the real meaning of the word sound, and in the magazine cover of the monster carrying off a woman in the torn dress, and Kant's declaration that space by its very nature is flat, and Hume's definition of a priori ideas as those "discoverable by the mere operation of thought, without dependence on what is anywhere existent in the universe"...

(Incidentally, I once read an SF story about a human male who entered into a sexual relationship with a sentient alien plant of appropriately squishy fronds; discovered that it was an androecious (male) plant; agonized about this for a bit; and finally decided that it didn't really matter at that point. And in Foglio and Pollotta's *Illegal Aliens,* the humans land on a planet inhabited by sentient insects, and see a movie advertisement showing a human carrying off a bug in a delicate chiffon dress. Just thought I'd mention that.)

Probability is in the Mind

Yesterday I spoke of the Mind Projection Fallacy, giving the example of the alien monster who carries off a girl in a torn dress for intended ravishing - a mistake which I imputed to the artist's tendency to think that a woman's sexiness is a property of the woman herself, `woman.sexiness`, rather than something that exists in the mind of an observer, and probably wouldn't exist in an alien mind.

The term "Mind Projection Fallacy" was coined by the late great Bayesian Master, E. T. Jaynes, as part of his long and hard-fought battle against the accursĂ¨d frequentists. Jaynes was of the opinion that probabilities were in the mind, not in the environment - that probabilities express ignorance, states of partial information; and if I am ignorant of a phenomenon, that is a fact about my state of mind, not a fact about the phenomenon.

I cannot do justice to this ancient war in a few words - but the classic example of the argument runs thus:

You have a coin.
The coin is biased.
You don't know which way it's biased or how much it's biased. Someone just told you, "The coin is biased" and that's all they said.
This is all the information you have, and the only information you have.

You draw the coin forth, flip it, and slap it down.

Now - before you remove your hand and look at the result - are you willing to say that you assign a 0.5 probability to the coin having come up heads?

The frequentist says, "No. Saying 'probability 0.5' means that the coin has an inherent propensity to come up heads as often as tails, so that if we flipped the coin infinitely many times, the ratio of heads to tails would approach 1:1. But we know that the coin is biased, so it can have any probability of coming up heads *except* 0.5."

The Bayesian says, "Uncertainty exists in the map, not in the territory. In the real world, the coin has either come up heads, or come up tails. Any talk of 'probability' must refer to the *information* that I have about the coin - my state of partial ignorance and partial knowledge - not just the coin itself. Furthermore, I have all sorts of theorems showing that if I don't treat my partial knowledge a certain way, I'll make stupid bets. If I've got to plan, I'll plan for a 50/50 state of uncertainty, where I don't weigh outcomes conditional on heads any more heavily in my mind than outcomes conditional on tails. You can call that number whatever you like, but it has to obey the probability laws on pain of stupidity. So I don't have the slightest hesitation about calling my outcome-weighting a probability."

I side with the Bayesians. You may have noticed that about me.

Even before a fair coin is tossed, the notion that it has an *inherent* 50% probability of coming up heads may be just plain wrong. Maybe you're holding the coin in such a way that it's just about guaranteed to come up heads, or tails, given the force at which you flip it, and the air currents around you. But, if you don't know which way the coin is biased on this one occasion, so what?

I believe there was a lawsuit where someone alleged that the draft lottery was unfair, because the slips with names on them were not being mixed thoroughly enough; and the judge replied, "To whom is it unfair?"

To make the coinflip experiment repeatable, as frequentists are wont to demand, we could build an automated coinflipper, and verify that the results were 50% heads and 50% tails. But maybe a robot with extra-sensitive eyes and a good grasp of physics, watching the autoflipper prepare to flip, could predict the coin's fall in advance - not with certainty, but with 90% accuracy. Then what would the *real* probability be?

There is no "real probability". The robot has one state of partial information. You have a different state of partial information. The coin itself has no mind, and doesn't assign a probability to anything; it just flips into the air, rotates a few times, bounces off some air molecules, and lands either heads or tails.

So that is the Bayesian view of things, and I would now like to point out a couple of classic brainteasers that derive their brain-*teasing* ability from the tendency to think of probabilities as inherent properties of objects.

Let's take the old classic: You meet a mathematician on the street, and she happens to mention that she has given birth to two children on two separate occasions. You ask: "Is at least one of your children a boy?" The mathematician says, "Yes, he is."

What is the probability that she has two boys? If you assume that the prior probability of a child being a boy is 1/2, then the probability that she has two boys, on the information given, is 1/3. The prior probabilities were: 1/4 two boys, 1/2 one boy one girl, 1/4 two girls. The mathematician's "Yes" response has probability $\sim$1 in the first two cases, and probability $\sim$0 in the third. Renormalizing leaves us with a 1/3 probability of two boys, and a 2/3 probability of one boy one girl.

But suppose that instead you had asked, "Is your eldest child a boy?" and the mathematician had answered "Yes." Then the probability of the mathematician having two boys would be 1/2. Since the eldest child is a boy, and the younger child can be anything it pleases.

Likewise if you'd asked "Is your youngest child a boy?" The probability of their being both boys would, again, be 1/2.

Now, if at least one child is a boy, it must be either the oldest child who is a boy, or the youngest child who is a boy. So how can the answer in the first case be different from the answer in the latter two?

Or here's a very similar problem: Let's say I have four cards, the ace of hearts, the ace of spades, the two of hearts, and the two of spades. I draw two cards at random. You ask me, "Are you holding at least one ace?" and I reply "Yes." What is the probability that I am holding a pair of aces? It is 1/5. There are six possible combinations of two cards, with equal prior probability, and you have just eliminated the possibility that I am holding a pair of twos. Of the five remaining combinations, only one combination is a pair of aces. So 1/5.

Now suppose that instead you asked me, "Are you holding the ace of spades?" If I reply "Yes", the probability that the other card is the ace of hearts is 1/3. (You know I'm holding the ace of spades, and there are three possibilities for the other card, only one of which is the ace of hearts.) Likewise, if you ask me "Are you holding the ace of hearts?" and I reply "Yes", the probability I'm holding a pair of aces is 1/3.

But then how can it be that if you ask me, "Are you holding at least one ace?" and I say "Yes", the probability I have a pair is 1/5? Either I must be holding the ace of spades or the ace of hearts, as you know; and either way, the probability that I'm holding a pair of aces is 1/3.

How can this be? Have I miscalculated one or more of these probabilities?

If you want to figure it out for yourself, do so now, because I'm about to reveal...

That all stated calculations are correct.

As for the paradox, there isn't one. The *appearance*of paradox comes from thinking that the probabilities must be properties of the cards themselves. The ace I'm holding has to be either hearts or spades; but that doesn't mean that your *knowledge about* my cards must be the same as if you *knew* I was holding hearts, or *knew* I was holding spades.

It may help to think of Bayes's Theorem:

$$P(H|E) = P(E|H)P(H) \ / \ P(E)$$

That last term, where you divide by P(E), is the part where you throw out all the possibilities that have been eliminated, and renormalize your probabilities over what remains.

Now let's say that you ask me, "Are you holding at least one ace?" *Before* I answer, your probability that I say "Yes" should be 5/6.

But if you ask me "Are you holding the ace of spades?", your prior probability that I say "Yes" is just 1/2.

So right away you can see that you're *learning* something very different in the two cases. You're going to be eliminating some different possibilities, and renormalizing using a different P(E). If you learn two different items of evidence, you shouldn't be surprised at ending up in two different states of partial information.

Similarly, if I ask the mathematician, "Is at least one of your two children a boy?" I expect to hear "Yes" with probability 3/4, but if I ask "Is your eldest child a boy?" I expect to hear "Yes" with probability 1/2. So it shouldn't be surprising that I end up in a different state of partial knowledge, depending on which of the two questions I ask.

The only reason for seeing a "paradox" is thinking as though the probability of holding a pair of aces is *a property of cards* that have at least one ace, or a property *of cards* that happen to contain the ace of spades. In which case, it would be paradoxical for card-sets containing at least one ace to have an inherent pair-probability of 1/5, while card-sets containing the ace of spades had an inherent pair-probability of 1/3, and card-sets containing the ace of hearts had an inherent pair-probability of 1/3.

Similarly, if you think a 1/3 probability of being both boys is an *inherent property* of child-sets that include at least one boy, then that is not consistent with child-sets of which the eldest is male having an *inherent* probability of 1/2 of being both boys, and child-sets of which the youngest is male having an inherent 1/2 probability of being both boys. It would be like saying, "All green apples weigh a pound, and all red apples weigh a pound, and all apples that are green or red weigh half a pound."

That's what happens when you start thinking as if probabilities are *in* things, rather than probabilities being states of partial information *about*things.

Probabilities express uncertainty, and it is only agents who can be uncertain. A blank map does not correspond to a blank territory. Ignorance is in the mind.

The Quotation is not the Referent

In classical logic, the operational definition of identity is that whenever 'A=B' is a theorem, you can substitute 'A' for 'B' in any theorem where B appears. For example, if `(2 + 2) = 4` is a theorem, and `((2 + 2) + 3) = 7` is a theorem, then `(4 + 3) = 7` is a theorem.

This leads to a problem which is usually phrased in the following terms: The morning star and the evening star happen to be the same object, the planet Venus. Suppose John knows that the morning star and evening star are the same object. Mary, however, believes that the morning star is the god Lucifer, but the evening star is the god Venus. John believes Mary believes that the morning star is Lucifer. Must John therefore (by substitution) believe that Mary believes that the evening star is Lucifer?

Or here's an even simpler version of the problem. $2 + 2 = 4$ is true; it is a theorem that `(((2 + 2) = 4) = TRUE)`. Fermat's Last Theorem is also true. So: I believe $2 + 2 = 4$ => I believe TRUE => I believe Fermat's Last Theorem.

Yes, I know this seems *obviously* wrong. But imagine someone writing a logical reasoning program using the principle "equal terms can always be substituted", and this happening to them. Now imagine them writing a paper about how to prevent it from happening. Now imagine someone else disagreeing with their solution. The argument is still going on.

P'rsnally, I would say that John is committing a type error, like trying to subtract 5 grams from 20 meters. "The morning star" is not the same *type* as the morning star, let alone the same thing. Beliefs are not planets.

> morning star = evening star
> "morning star" $\neq$ "evening star"

The problem, in my view, stems from the failure to enforce the type distinction between beliefs and things. The original error was writing an AI that stores its beliefs about Mary's beliefs about "the morning star" using the same representation as in its beliefs about the morning star.

If Mary believes the "morning star" is Lucifer, that doesn't mean Mary believes the "evening star" is Lucifer, because "morning star" $\neq$ "evening star". The whole paradox stems from the failure to use quote marks in appropriate places.

You may recall that this is not the first time I've talked about enforcing type discipline - the last time was when I spoke about the error of confusing expected utilities with utilities. It is immensely helpful, when one is first learning physics,

to learn to keep track of one's units - it may seem like a bother to keep writing down 'cm' and 'kg' and so on, until you notice that (a) your answer seems to be the wrong order of magnitude and (b) it is expressed in seconds per square gram.

Similarly, beliefs are different things than planets. If we're talking about human beliefs, at least, then: Beliefs live in brains, planets live in space. Beliefs weigh a few micrograms, planets weigh a lot more. Planets are larger than beliefs... but you get the idea.

Merely putting quote marks around "morning star" seems insufficient to prevent people from confusing it with the morning star, due to the visual similarity of the text. So perhaps a better way to enforce type discipline would be with a visibly different encoding:

> morning star = evening star
> 13.15.18.14.9.14.7.0.19.20.1.18 ≠ 5.22.5.14.9.14.7.0.19.20.1.18

Studying mathematical logic may also help you learn to distinguish the quote and the referent. In mathematical logic, `|- P` (P is a theorem) and `|- []'P'` (it is provable that there exists an encoded proof of the encoded sentence P in some encoded proof system) are very distinct propositions. If you drop a level of quotation in mathematical logic, it's like dropping a metric unit in physics - you can derive visibly ridiculous results, like "The speed of light is 299,792,458 meters long."

Alfred Tarski once tried to define the meaning of 'true' using an infinite family of sentences:

> ("Snow is white" is true) if and only (snow is white)
> ("Weasels are green" is true) if and only if (weasels are green)
> ...

When sentences like these start seeming meaningful, you'll know that you've started to distinguish between encoded sentences and states of the outside world.

Similarly, the notion of truth is quite different from the notion of *reality*. Saying "true" *compares* a belief to reality. Reality itself does not need to be compared to any beliefs in order to be real. Remember this the next time someone claims that nothing is true.

Qualitatively Confused

I suggest that a primary cause of confusion about the distinction between "belief", "truth", and "reality" is qualitative thinking about beliefs.

Consider the archetypal postmodernist attempt to be clever:

> "The Sun goes around the Earth" is true for Hunga Huntergatherer, but "The Earth goes around the Sun" is true for Amara Astronomer! Different societies have different truths!

No, different societies have different *beliefs.* Belief is of a different type than truth; it's like comparing apples and probabilities.

> Ah, but there's no difference between the way you use the word 'belief' and the way you use the word 'truth'! Whether you say, "I believe 'snow is white' ", or you say, " 'Snow is white' is true", you're expressing exactly the same opinion.

No, these sentences mean quite different things, which is how I can *conceive* of the possibility that my beliefs are false.

> Oh, you claim to *conceive* it, but you never *believe* it. As Wittgenstein said, "If there were a verb meaning 'to believe falsely', it would not have any significant first person, present indicative."

And that's what I mean by putting my finger on qualitative reasoning as the source of the problem. The dichotomy between belief and disbelief, being binary, is confusingly similar to the dichotomy between truth and untruth.

So let's use quantitative reasoning instead. Suppose that I assign a 70% probability to the proposition that snow is white. It follows that I think there's around a 70% chance that the sentence "snow is white" will turn out to be true. If the sentence "snow is white" is true, is my 70% probability assignment to the proposition, also "true"? Well, it's more true than it would have been if I'd assigned 60% probability, but not so true as if I'd assigned 80% probability.

When talking about the correspondence between a probability assignment and reality, a better word than "truth" would be "accuracy". "Accuracy" sounds more quantitative, like an archer shooting an arrow: how close did your probability assignment strike to the center of the target?

To make a long story short, it turns out that there's a very natural way of scoring the accuracy of a probability assignment, as compared to reality: just take the logarithm of the probability assigned to the real state of affairs.

So if snow is white, my belief "70%: 'snow is white' " will score –0.51 bits: $\text{Log}_2(0.7) = -0.51$.

But what if snow is not white, as I have conceded a 30% probability is the case? If "snow is white" is false, my belief "30% probability: 'snow is not white' " will score –1.73 bits. Note that $-1.73 < -0.51$, so I have done worse.

About how accurate do I think my own beliefs are? Well, my expectation over the score is 70% * –0.51 + 30% * –1.73 = –0.88 bits. If snow is white, then

my beliefs will be more accurate than I expected; and if snow is not white, my beliefs will be less accurate than I expected; but in neither case will my belief be *exactly* as accurate as I expected on average.

All this should not be confused with the statement "I assign 70% credence that 'snow is white'." I may well believe *that* proposition with probability ∼1 - be quite certain that this is in fact my belief. If so I'll expect my meta-belief "∼1: 'I assign 70% credence that "snow is white" '" to score ∼0 bits of accuracy, which is as good as it gets.

Just because I am uncertain about snow, does not mean I am uncertain about my *quoted probabilistic beliefs*. Snow is out there, my beliefs are inside me. I may be a great deal less uncertain about how uncertain I am about snow, than I am uncertain about snow. (Though beliefs about beliefs are not always accurate.)

Contrast this probabilistic situation to the qualitative reasoning where I just believe that snow is white, and believe that I believe that snow is white, and believe "'snow is white' is true", and believe "my belief ' "snow is white" is true' is correct", etc. Since all the quantities involved are 1, it's easy to mix them up.

Yet the nice distinctions of quantitative reasoning will be short-circuited if you start thinking "' "snow is white" with 70% probability' is *true*", which is a type error. It is a true fact about you, that you *believe* "70% probability: 'snow is white'"; but that does not mean the probability assignment *itself* can possibly be "true". The belief scores either –0.51 bits or –1.73 bits of accuracy, depending on the actual state of reality.

The cognoscenti will recognize "' "snow is white" with 70% probability' is true" as the mistake of thinking that probabilities are inherent properties of things.

From the inside, our beliefs about the world look like the world, and our beliefs about our beliefs look like beliefs. When you see the world, you are experiencing a belief from the inside. When you notice yourself believing something, you are experiencing a belief about belief from the inside. So if your internal representations of belief, and belief about belief, are dissimilar, then you are less likely to mix them up and commit the Mind Projection Fallacy - I hope.

When you think in probabilities, your beliefs, and your beliefs about your beliefs, will hopefully not be represented similarly enough that you mix up belief and accuracy, or mix up accuracy and reality. When you think in probabilities *about the world*, your beliefs will be represented with probabilities ∈ (0, 1). Unlike the truth-values of propositions, which are in {true, false}. As for the accuracy of your probabilistic belief, you can represent that in the range (-∞, 0). Your probabilities *about your beliefs* will typically be extreme. And things themselves - why, they're just red, or blue, or weighing 20 pounds, or whatever.

Thus we will be less likely, perhaps, to mix up the map with the territory.

This type distinction may also help us remember that *uncertainty* is a state of mind. A coin is not *inherently* 50% uncertain of which way it will land. The coin is not a belief processor, and does not have partial information about itself. In qualitative reasoning you can create a belief that corresponds very straightforwardly to the coin, like "The coin will land heads". This belief will be true or false *depending on* the coin, and there will be a transparent implication from the truth or falsity of the belief, to the facing side of the coin.

But even under qualitative reasoning, to say that the coin *itself* is "true" or "false" would be a severe type error. The coin is not a belief, it is a coin. The territory is not the map.

If a coin cannot be true or false, how much less can it assign a 50% probability to itself?

Reductionism

Almost one year ago, in April 2007, Matthew C submitted the following suggestion for an Overcoming Bias topic:

> "How and why the current reigning philosophical hegemon (reductionistic materialism) is obviously correct [...], while the reigning philosophical viewpoints of all past societies and civilizations are obviously suspect -"

I remember this, because I looked at the request and deemed it legitimate, but I knew I couldn't do that topic until I'd started on the Mind Projection Fallacy sequence, which wouldn't be for a while...

But now it's time to begin addressing this question. And while I haven't yet come to the "materialism" issue, we can now start on "reductionism".

First, let it be said that I do indeed hold that "reductionism", according to the meaning I will give for that word, is obviously correct; and to perdition with any past civilizations that disagreed.

This seems like a strong statement, at least the first part of it. General Relativity seems well-supported, yet who knows but that some future physicist may overturn it?

On the other hand, we are never going *back*to Newtonian mechanics. The ratchet of science turns, but it does not turn in reverse. There are cases in scientific history where a theory suffered a wound or two, and then bounced back; but when a theory takes as many arrows through the chest as Newtonian mechanics, it *stays dead.*

"To hell with what past civilizations thought" seems safe enough, when past civilizations believed in something that has been falsified to the trash heap of history.

And reductionism is not so much a positive hypothesis, as the *absence* of belief - in particular, disbelief in a form of the Mind Projection Fallacy.

I once met a fellow who claimed that he had experience as a Navy gunner, and he said, "When you fire artillery shells, you've got to compute the trajectories using Newtonian mechanics. If you compute the trajectories using relativity, you'll get the wrong answer."

And I, and another person who was present, said flatly, "No." I added, "You might not be able to compute the trajectories fast enough to get the answers in time - maybe that's what you mean? But the relativistic answer will always be more accurate than the Newtonian one."

"No," he said, "I mean that relativity will give you the *wrong answer,* because things moving at the speed of artillery shells are governed by Newtonian mechanics, not relativity."

"If that were really true," I replied, "you could publish it in a physics journal and collect your Nobel Prize."* *

Standard physics uses the same *fundamental* theory to describe the flight of a Boeing 747 airplane, and collisions in the Relativistic Heavy Ion Collider. Nuclei and airplanes alike, according to our understanding, are obeying special relativity, quantum mechanics, and chromodynamics.

But we use entirely different *models* to understand the aerodynamics of a 747 and a collision between gold nuclei in the RHIC. A computer modeling the aerodynamics of a 747 may not contain a single token, a single bit of RAM, that represents a quark.

So is the 747 made of something other than quarks? No, you're just *modeling* it with *representational elements* that do not have a one-to-one correspondence with the quarks of the 747. The map is not the territory.

Why *not* model the 747 with a chromodynamic representation? Because then it would take a gazillion years to get any answers out of the model. Also we could not store the model on all the memory on all the computers in the world, as of 2008.

As the saying goes, "The map is not the territory, but you can't fold up the territory and put it in your glove compartment." Sometimes you need a smaller map to fit in a more cramped glove compartment - but this does not change the territory. The scale of a map is not a fact about the territory, it's a fact about the map.

If it *were* possible to build and run a chromodynamic model of the 747, it would yield accurate predictions. Better predictions than the aerodynamic model, in fact.

To build a fully accurate model of the 747, it is not necessary, in principle, for the model to contain explicit descriptions of things like airflow and lift. There does not have to be a single token, a single bit of RAM, that corresponds to the position of the wings. It is possible, in principle, to build an accurate model of the 747 that makes no mention of anything *except* elementary particle fields and fundamental forces.

"What?" cries the antireductionist. "Are you telling me the 747 *doesn't really have wings?* I can see the wings right there!"

The notion here is a subtle one. It's not *just* the notion that an object can have different descriptions at different levels.

It's the notion that "having different descriptions at different levels" is *itself* something you say that belongs in the realm of Talking About Maps, not the realm of Talking About Territory.

It's not that the *airplane itself,* the *laws of physics themselves,* use different descriptions at different levels - as yonder artillery gunner thought. Rather *we,* for our convenience, use different simplified models at different levels.

If you looked at the ultimate chromodynamic model, the one that contained only elementary particle fields and fundamental forces, that model would contain all the facts about airflow and lift and wing positions - but these facts would be implicit, rather than explicit.

You, looking *at* the model, and thinking *about* the model, would be able to figure out where the wings were. Having figured it out, there would be an explicit representation in your mind of the wing position - an explicit computational object, there in your neural RAM. *In your mind.*

You might, indeed, deduce all sorts of explicit descriptions of the airplane, at various levels, and even explicit rules for how your models at different levels interacted with each other to produce combined predictions -

And the way that algorithm feels from inside, is that the airplane would *seem* to be made up of many levels at once, interacting with each other.

The way a belief *feels from inside*, is that you seem to be looking straight at reality. When it actually *seems* that you're looking at a belief, as such, you are really experiencing a belief about belief.

So when your mind simultaneously believes explicit descriptions of many different levels, and believes explicit rules for transiting between levels, as part of an efficient combined model, it *feels like* you are seeing a system that is *made of* different level descriptions and their rules for interaction.

But this is just the brain trying to be efficiently compress an object that it cannot remotely begin to model on a fundamental level. The airplane is too large. Even a hydrogen atom would be too large. Quark-to-quark interactions are insanely intractable. You can't handle the*truth.*

But the way physics *really* works, as far as we can tell, is that there is *only* the most basic level - the elementary particle fields and fundamental forces. You can't handle the raw truth, but reality can handle it without the slightest simplification. (I wish I knew where Reality got its computing power.)

The laws of physics do not contain distinct additional causal entities that correspond to lift or airplane wings, the way that *the mind of an engineer* contains distinct additional *cognitive* entities that correspond to lift or airplane wings.

This, as I see it, is the thesis of reductionism. Reductionism is not a positive belief, but rather, a disbelief that the higher levels of simplified multilevel models are out there in the territory. Understanding this on a gut level dissolves the question of "How can you say the airplane doesn't really have wings, when I can see the wings right there?" The critical words are *really* and *see*.

Explaining vs Explaining Away

John Keats's *Lamia* (1819) surely deserves some kind of award for Most Famously Annoying Poetry:

> . . . Do not all charms fly
> At the mere touch of cold philosophy?
> There was an awful rainbow once in heaven:
> We know her woof, her texture; she is given
> In the dull catalogue of common things.
> Philosophy will clip an Angel's wings,
> Conquer all mysteries by rule and line,
> Empty the haunted air, and gnomed mine -
> Unweave a rainbow. . .

My usual reply ends with the phrase: "If we cannot learn to take joy in the merely real, our lives will be empty indeed." I shall expand on that tomorrow.

Today I have a different point in mind. Let's just take the lines:

> Empty the haunted air, and gnomed mine -
> Unweave a rainbow. . .

Apparently "the mere touch of cold philosophy", i.e., the truth, has destroyed:

- Haunts in the air
- Gnomes in the mine
- Rainbows

Which calls to mind a rather different bit of verse:

> One of these things
> Is not like the others
> One of these things
> Doesn't belong

The air has been emptied of its haunts, and the mine de-gnomed - but the rainbow is still there!

In "Righting a Wrong Question", I wrote:

> Tracing back the chain of causality, step by step, I discover that my belief that I'm wearing socks is fully explained by the fact that I'm wearing socks... On the other hand, if I see a mirage of a lake in the desert, the correct causal explanation of my vision does not involve the fact of any actual lake in the desert. In this case, my belief in the lake is not just *explained,* but *explained away.*

The rainbow was *explained.* The haunts in the air, and gnomes in the mine, were *explained away.*

I think this is the key distinction that anti-reductionists don't get about reductionism.

You can see this failure to get the distinction in the classic objection to reductionism:

> If reductionism is correct, then even your belief in reductionism is just the mere result of the motion of molecules - why should I listen to anything you say?

The key word, in the above, is *mere;* a word which implies that accepting reductionism would explain*away* all the reasoning processes leading up to my acceptance of reductionism, the way that an optical illusion is explained *away.*

But you can explain how a cognitive process works without it being "mere"! My belief that I'm wearing socks is a mere result of my visual cortex reconstructing nerve impulses sent from my retina which received photons reflected off my socks... which is to say, according to scientific reductionism, my belief that I'm wearing socks is a mere result of the fact that I'm wearing socks.

What could be going on in the anti-reductionists' minds, such that they would put rainbows and belief-in-reductionism, in the same category as haunts and gnomes?

Several things are going on simultaneously. But for now let's focus on the basic idea introduced yesterday: The Mind Projection Fallacy between a multi-level map and a mono-level territory.

(I.e: There's no way you can model a 747 quark-by-quark, so you've *got* to use a multi-level map with explicit cognitive representations of wings, airflow, and so on. This doesn't mean there's a multi-level territory. The true laws of physics, to the best of our knowledge, are only over elementary particle fields.)

I think that when physicists say "There are no *fundamental* rainbows," the anti-reductionists hear, "There are no rainbows."

If you don't distinguish between the multi-level map and the mono-level territory, then when someone tries to explain to you that the rainbow is not a fundamental thing in physics, acceptance of this will *feel like* erasing rainbows from your multi-level map, which *feels like* erasing rainbows from the world.

When Science says "tigers are not *elementary* particles, they are made of quarks" the anti-reductionist hears this as the same sort of dismissal as "we looked in your garage for a dragon, but there was just empty air".

What scientists did to rainbows, and what scientists did to gnomes, seemingly felt the same to Keats...

In support of this sub-thesis, I deliberately used several phrasings, in my discussion of Keats's poem, that were Mind Projection Fallacious. If you didn't notice, this would seem to argue that such fallacies are customary enough to pass unremarked.

For example:

> "The air has been emptied of its haunts, and the mine de-gnomed - but the rainbow is still there!"

Actually, Science emptied the *model of* air of *belief in* haunts, and emptied the *map of* the mine of *representations of* gnomes. Science did not actually - as Keats's poem itself would have it - take real Angel's wings, and destroy them with a cold touch of truth. In reality there *never were* any haunts in the air, or gnomes in the mine.

Another example:

> "What scientists did to rainbows, and what scientists did to gnomes, seemingly felt the same to Keats."

Scientists didn't *do* anything *to* gnomes, only to "gnomes". The quotation is not the referent.

But if you commit the Mind Projection Fallacy - and by default, our beliefs just feel like the way the world *is* - then at time T=0, the mines (apparently)

contain gnomes; at time T=1 a scientist dances across the scene, and at time T=2 the mines (apparently) are empty. Clearly, there used to be gnomes there, but the scientist killed them.

Bad scientist! No poems for you, gnomekiller!

Well, that's how it *feels,* if you get emotionally attached to the gnomes, and then a scientist says there aren't any gnomes. It takes a strong mind, a deep honesty, and a deliberate effort to say, at this point, "That which can be destroyed by the truth should be," and "The scientist hasn't taken the gnomes away, only taken my delusion away," and "I never held just title to my belief in gnomes in the first place; I have not been deprived of anything I *rightfully* owned," and "If there are gnomes, I desire to believe there are gnomes; if there are no gnomes, I desire to believe there are no gnomes; let me not become attached to beliefs I may not want," and all the other things that rationalists are supposed to say on such occasions.

But with the rainbow it is not even necessary to go that far. The rainbow is *still there!*

Fake Reductionism

> There was an awful rainbow once in heaven:
> We know her woof, her texture; she is given
> In the dull catalogue of common things.
> — John Keats, *Lamia*

I am guessing - though it is only a guess - that Keats himself did *not* know the woof and texture of the rainbow. Not the way that Newton understood rainbows. Perhaps not even at all. Maybe Keats just read, somewhere, that Newton had explained the rainbow as "light reflected from raindrops" -

- which was actually known in the 13th century. Newton only added a refinement by showing that the light was decomposed into colored parts, rather than transformed in color. But that put rainbows back in the news headlines. And so Keats, with Charles Lamb and William Wordsworth and Benjamin Haydon, drank "Confusion to the memory of Newton" because "he destroyed the poetry of the rainbow by reducing it to a prism." That's one reason to suspect Keats didn't understand the subject too deeply.

I am guessing, though it is only a guess, that Keats could *not* have sketched out on paper why rainbows only appear when the Sun is behind your head, or why the rainbow is an arc of a circle.

If so, Keats had a Fake Explanation. In this case, a *fake reduction*. He'd been *told that* the rainbow had been reduced, but it had not actually *been reduced* in his model of the world.

This is another of those distinctions that anti-reductionists fail to get - the difference between professing the flat fact that something is reducible, and *seeing* it.

In this, the anti-reductionists are not too greatly to be blamed, for it is part of a general problem.

I've written before on seeming knowledge that is not knowledge, and beliefs that are not *about* their supposed objects but only recordings to recite back in the classroom, and words that operate as stop signs for curiosity rather than answers, and technobabble which only conveys membership in the literary genre of "science"...

There is a very great distinction between being able to *see* where the rainbow comes from, and playing around with prisms to confirm it, and maybe making a rainbow yourself by spraying water droplets -

- versus some dour-faced philosopher just *telling* you, "No, there's nothing special about the rainbow. Didn't you hear? Scientists have explained it away. Just something to do with raindrops or whatever. Nothing to be excited about."

I think this distinction probably accounts for a hell of a lot of the deadly existential emptiness that supposedly accompanies scientific reductionism.

You have to interpret the anti-reductionists' experience of "reductionism", not in terms of their *actually seeing* how rainbows work, not in terms of their having the critical "Aha!", but in terms of their being told that the password is "Science". The effect is just to move rainbows to a different *literary genre* - a literary genre they have been taught to regard as boring.

For them, the effect of hearing "Science has explained rainbows!" is to hang up a sign over rainbows saying, "This phenomenon has been labeled BORING by order of the Council of Sophisticated Literary Critics. Move along."

And that's all the sign says: only that, and nothing more.

So the literary critics have their gnomes yanked out by force; not dissolved in insight, but removed by flat order of authority. They are given no beauty to replace the hauntless air, no genuine understanding that could be interesting in its own right. Just a label saying, "Ha! You thought rainbows were pretty? You poor, unsophisticated fool. This is part of the literary genre of science, of dry and solemn incomprehensible words."

That's how anti-reductionists experience "reductionism".

Well, can't blame Keats, poor lad probably wasn't raised right.

But he dared to drink "Confusion to the memory of Newton"?* *

I propose "To the memory of Keats's confusion" as a toast for rationalists. Cheers.

Savanna Poets

> "Poets say science takes away from the beauty of the stars - mere globs of gas atoms. Nothing is"mere". I too can see the stars on a desert night, and feel them. But do I see less or more?
> "The vastness of the heavens stretches my imagination - stuck on this carousel my little eye can catch one-million-year-old light. A vast pattern - of which I am a part - perhaps my stuff was belched from some forgotten star, as one is belching there. Or see them with the greater eye of Palomar, rushing all apart from some common starting point when they were perhaps all together. What is the pattern, or the meaning, or the why? It does not do harm to the mystery to know a little about it.
> "For far more marvelous is the truth than any artists of the past imagined! Why do the poets of the present not speak of it?
> "What men are poets who can speak of Jupiter if he were like a man, but if he is an immense spinning sphere of methane and ammonia must be silent?"
> — Richard Feynman, *The Feynman Lectures on Physics,*Vol I, p. 3–6 (line breaks added)

That's a real question, there on the last line - what kind of poet can write about Jupiter the god, but not Jupiter the immense sphere? Whether or not Feynman meant the question rhetorically, it has a real answer:

If Jupiter is like us, he can fall in love, and lose love, and regain love.
If Jupiter is like us, he can strive, and rise, and be cast down.
If Jupiter is like us, he can laugh or weep or dance.

If Jupiter is an immense spinning sphere of methane and ammonia, it is more difficult for the poet to make us feel.

There are poets and storytellers who say that the Great Stories are timeless, and they never change, they only ever retold. They say, with pride, that Shakespeare and Sophocles are bound by ties of craft stronger than mere centuries; that the two playwrights could have swapped times without a jolt.

Donald Brown once compiled a list of over two hundred "human universals", found in all (or a vast supermajority of) studied human cultures, from San Francisco to the !Kung of the Kalahari Desert. Marriage is on the list, and incest avoidance, and motherly love, and sibling rivalry, and music and envy and dance and storytelling and aesthetics, and ritual magic to heal the sick, and poetry in spoken lines separated by pauses -

No one who knows anything about evolutionary psychology could be expected to deny it: The strongest emotions we have are deeply engraved, blood and bone, brain and DNA.

It might take a bit of tweaking, but you probably *could* tell "Hamlet" sitting around a campfire on the ancestral savanna.

So one can see why John "Unweave a rainbow" Keats might feel something had been lost, on being told that the rainbow was sunlight scattered from raindrops. Raindrops don't dance.

In the Old Testament, it is written that God once destroyed the world with a flood that covered all the land, drowning all the horribly guilty men and women of the world along with their horribly guilty babies, but Noah built a gigantic wooden ark, etc., and after most of the human species was wiped out, God put rainbows in the sky as a sign that he wouldn't do it again. At least not with water.

You can see how Keats would be *shocked* that this beautiful story was contradicted by modern science. Especially if (as I described yesterday) Keats had no real understanding of rainbows, no "Aha!" insight that could be fascinating in its own right, to replace the drama subtracted -

Ah, but maybe Keats would be right to be disappointed *even if* he knew the math. The Biblical story of the rainbow is a tale of bloodthirsty murder and smiling insanity. How could anything about raindrops and refraction properly replace that? Raindrops don't scream when they die.

So science takes the romance away (says the Romantic poet), and what you are given back, never matches the drama of the original -

(that is, the original delusion)

- even if you do know the equations, because the equations are not about strong emotions.

That is the strongest rejoinder I can think of, that any Romantic poet could have said to Feynman - though I can't remember ever hearing it said.

You can guess that I don't agree with the Romantic poets. So my own stance is this:

It is not *necessary* for Jupiter to be like a human, because *humans*are like humans. If Jupiter is an immense spinning sphere of methane and ammonia, that doesn't mean that love and hate are emptied from the universe. There *are* still loving and hating minds in the universe. *Us.*

With more than six billion of us at the last count, does Jupiter really need to be on the list of potential protagonists?

It is not *necessary* to tell the Great Stories about planets or rainbows. They play out all over our world, every day. Every day, someone kills for revenge; every day, someone kills a friend by mistake; every day, upward of a hundred thousand people fall in love. And even if this were not so, you could write fiction about humans - not about Jupiter.

Earth is old, and has played out the same stories many times beneath the Sun. I do wonder if it might not be time for some of the Great Stories to change. For me, at least, the story called "Goodbye" has lost its charm.

The Great Stories are not timeless, because the human species is not timeless. Go far enough back in hominid evolution, and no one will understand *Hamlet*. Go far enough back in time, and you won't find any brains.

The Great Stories are not eternal, because the human species, *Homo sapiens sapiens,* is not eternal. I most sincerely doubt that we have another thousand years to go in our current form. I do not say this in sadness: I think we can do better.

I would not like to see all the Great Stories lost completely, in our future. I see very little difference between that outcome, and the Sun falling into a black hole.

But the Great Stories in their current forms have *already been* told, over and over. I do not think it ill if some of them should change their forms, or diversify their endings.

"And they lived happily ever after" seems worth trying at least once.

The Great Stories can and should diversify, as humankind grows up. Part of that ethic is the idea that when we find strangeness, we should respect it enough to tell its story truly. Even if it makes writing poetry a little more difficult.

If you are a good enough poet to write an ode to an immense spinning sphere of methane and ammonia, you are writing something *original,* about a newly discovered part of the real universe. It may not be as dramatic, or as gripping, as Hamlet. But the tale of Hamlet has already been told! If you write of Jupiter as though it were a human, then you are making our map of the universe just a little more impoverished of complexity; you are forcing Jupiter into the mold of all the stories that have already been told of Earth.

James Thomson's "A Poem Sacred to the Memory of Sir Isaac Newton", which praises the rainbow for what it *really* is - you can argue whether or not Thomson's poem is as gripping as John Keats's Lamia who was loved and lost. But tales of love and loss and cynicism had *already been* told, far away in ancient Greece, and no doubt many times before. Until we understood the rainbow as a thing *different* from tales of human-shaped magic, the true story of the rainbow could not be poeticized.

The border between science fiction and space opera was once drawn as follows: If you can take the plot of a story and put it back in the Old West, or

the Middle Ages, without changing it, then it is not *real* science fiction. In real science fiction, the science is intrinsically part of the plot - you can't move the story from space to the savanna, not without losing something.

Richard Feynman asked: "What men are poets who can speak of Jupiter if he were like a man, but if he is an immense spinning sphere of methane and ammonia must be silent?"

They are *savanna poets,* who can *only* tell stories that would have made sense around a campfire ten thousand years ago. Savanna poets, who can tell *only* the Great Stories in their classic forms, and nothing more.

Joy in the Merely Real

> . . . Do not all charms fly
> At the mere touch of cold philosophy?
> There was an awful rainbow once in heaven:
> We know her woof, her texture; she is given
> In the dull catalogue of common things.
> — John Keats, *Lamia*
>
> "Nothing is 'mere'."
> — Richard Feynman* *

You've got to admire that phrase, "dull catalogue of common things". What is it, exactly, that goes in this catalogue? Besides rainbows, that is?

Why, things that are mundane, of course. Things that are normal; things that are unmagical; things that are known, or knowable; things that play by the rules (or that play by *any* rules, which makes them boring); things that are part of the ordinary universe; things that are, in a word, *real.*

Now that's what I call setting yourself up for a fall.

At that rate, sooner or later you're going to be disappointed in *everything* - either it will turn out not to exist, or even worse, it will turn out to be real.

If we cannot take joy in things that are merely real, our lives will *always* be empty.

For what sin are rainbows demoted to the dull catalogue of common things? For the sin of having a scientific explanation. "We know her woof, her texture", says Keats - an interesting use of the word "we", because I suspect that Keats didn't know the explanation himself. I suspect that just being told that someone else knew was too much for him to take. I suspect that just the notion of rainbows being scientifically explicable *in principle* would have been too much to take. And if Keats didn't think like that, well, I know plenty of people who do.

I have already remarked that nothing is *inherently* mysterious - nothing that actually exists, that is. If I am ignorant about a phenomenon, that is a fact about my state of mind, not a fact about the phenomenon; to worship a phenomenon because it seems so wonderfully mysterious, is to worship your own ignorance; a blank map does not correspond to a blank territory, it is just somewhere we haven't visited yet, etc. etc...

Which is to say that *everything* - everything that *actually*exists - is liable to end up in "the dull catalogue of common things", sooner or later.

Your choice is either:

- Decide that things are allowed to be unmagical, knowable, scientifically explicable, in a word, *real,* and yet still worth caring about;
- Or go about the rest of your life suffering from existential ennui that is *unresolvable.*

(Self-deception might be an option for others, but not for you.)

This puts quite a different complexion on the bizarre habit indulged by those strange folk called *scientists,* wherein they suddenly become fascinated by pocket lint or bird droppings or rainbows, or some other ordinary thing which world-weary and sophisticated folk would never give a second glance.

You might say that scientists - at least *some*scientists - are those folk who are *in principle* capable of enjoying life in the real universe.

Joy in Discovery

> "Newton was the greatest genius who ever lived, and the most fortunate; for we cannot find more than once a system of the world to establish."
> — Lagrange

I have more fun discovering things for myself than reading about them in textbooks. This is right and proper, and only to be expected.

But discovering something that *no one else knows* - being the *first* to unravel the secret -

There is a story that one of the first men to realize that stars were burning by fusion - plausible attributions I've seen are to Fritz Houtermans and Hans Bethe - was walking out with his girlfriend of a night, and she made a comment on how beautiful the stars were, and he replied: "Yes, and right now, I'm the only man in the world who knows why they shine."

It is attested by numerous sources that this experience, being the first person to solve a major mystery, is a *tremendous* high. It's probably the closest experience you can get to taking drugs, without taking drugs - though I wouldn't know.

That can't be healthy.

Not that I'm objecting to the euphoria. It's the exclusivity clause that bothers me. Why should a discovery be worth *less,* just because someone *else* already knows the answer?

The most charitable interpretation I can put on the psychology, is that you don't struggle with a single problem for months or years if it's something you can just look up in the library. And that the tremendous high comes from having hit the problem from every angle you can manage, and having bounced; and then having analyzed the problem again, using every idea you can think of, and all the data you can get your hands on - making progress a little at a time - so that when, *finally,* you crack through the problem, all the dangling pieces and unresolved questions fall into place at once, like solving a dozen locked-room murder mysteries with a single clue.

And more, the understanding you get is *real* understanding - understanding that embraces all the clues you studied to solve the problem, when you didn't yet know the answer. Understanding that comes from asking questions day after day and worrying at them; understanding that no one else can get (no matter how much you tell them the answer) unless they spend months studying the problem in its historical context, even after it's been solved - and even then, they won't get the high of solving it all at once.

That's one possible reason why James Clerk Maxwell might have had more fun *discovering* Maxwell's Equations, than you had fun reading about them.

A slightly less charitable reading is that the tremendous high comes from what is termed, in the *politesse* of social psychology, "commitment" and "consistency" and "cognitive dissonance"; the part where we value something more highly *just* because it took more work to get it. The studies showing that subjective fraternity pledges to a harsher initiation, causes them to be more convinced of the value of the fraternity - identical wine in higher-priced bottles being rated as tasting better - that sort of thing.

Of course, if you just have more fun solving a puzzle than being told its answer, because you enjoy doing the cognitive work for its own sake, there's nothing wrong with that. The less charitable reading would be if charging $100 to be told the answer to a puzzle, made you think the answer was more interesting, worthwhile, important, surprising, etc. than if you got the answer for free.

(I strongly suspect that a major part of science's PR problem in the population at large is people who instinctively believe that if knowledge is given away for free, it cannot be important. If you had to undergo a fearsome initiation ritual to be told the truth about evolution, maybe people would be more satisfied with the answer.)

The really uncharitable reading is that the joy of first discovery is about status. Competition. Scarcity. Beating everyone else to the punch. It doesn't matter whether you have a 3-room house or a 4-room house, what matters is having a bigger house than the Joneses. A 2-room house would be fine, if you could only ensure that the Joneses had even less.

I don't object to competition as a matter of principle. I don't think that the game of Go is barbaric and should be suppressed, even though it's zero-sum. But if the euphoric joy of scientific discovery *has* to be about scarcity, that means it's only available to one person per civilization for any given truth.

If the joy of scientific discovery is one-shot per discovery, then, from a fun-theoretic perspective, Newton probably used up a substantial increment of the total Physics Fun available over the entire history of Earth-originating intelligent life. That selfish bastard explained the orbits of planets *and* the tides.

And really the situation is even worse than this, because in the Standard Model of physics (discovered by bastards who spoiled the puzzle for everyone else) the universe is spatially infinite, inflationarily branching, and branching via decoherence, which is at least three different ways that Reality is exponentially or infinitely large

So aliens, or alternate Newtons, or just Tegmark duplicates of Newton, may all have discovered gravity before *our* Newton did - if you believe that "before" means anything relative to those kinds of separations.

When that thought first occurred to me, I actually found it quite uplifting. Once I realized that someone, somewhere in the expanses of space and time, already knows the answer to any answerable question - even biology questions and history questions; there are other decoherent Earths - then I realized how silly it was to think as if the joy of discovery ought to be limited to one person. It becomes a fully inescapable source of unresolvable existential angst, and I regard that as a reductio.

The consistent solution which maintains the *possibility* of fun, is to stop worrying about what other people know. If you don't know the answer, it's a mystery to you. If you can raise your hand, and clench your fingers into a fist, and you've got no idea of how your brain is doing it - or even what exact muscles lay beneath your skin - you've got to consider yourself just as ignorant as a hunter-gatherer. Sure, someone else knows the answer - but back in the hunter-gatherer days, someone else in an alternate Earth, or for that matter, someone else in the future, knew what the answer was. Mystery, and the joy of finding out, is either a personal thing, or it doesn't exist at all - and I prefer to say it's personal.

The joy of assisting your civilization by telling it something it doesn't already know, does tend to be one-shot per discovery per civilization; that kind of value is conserved, as are Nobel Prizes. And the prospect of that reward may be what it takes to keep you focused on one problem for the years required to

develop a really *deep* understanding; plus, working on a problem unknown to your civilization is a sure-fire way to avoid reading any spoilers.

But as part of my general project to undo this idea that rationalists have less fun, I want to restore the magic and mystery to every part of the world which you do not *personally* understand, regardless of what other knowledge may exist, far away in space and time, or even in your next-door neighbor's mind. If *you* don't know, it's a mystery. And now think of how many things you don't know! (If you can't think of anything, you have other problems.) Isn't the world suddenly a much more mysterious and magical and *interesting* place? As if you'd been transported into an alternate dimension, and had to learn all the rules from scratch?

> "A friend once told me that I look at the world as if I've never seen it before. I thought, that's a nice compliment... Wait! I never *have* seen it before! What — did everyone else get a preview?"
> — Ran Prieur

Bind Yourself to Reality

So perhaps you're reading all this, and asking: "Yes, but what does this have to do with reductionism?"

Partially, it's a matter of leaving a line of retreat. It's not easy to take something *important* apart into components, when you're convinced that this removes magic from the world, unweaves the rainbow. I do plan to take certain things apart, on this blog; and I prefer not to create pointless existential anguish.

Partially, it's the crusade against Hollywood Rationality, the concept that understanding the rainbow subtracts its beauty. The rainbow is still beautiful *plus* you get the beauty of physics.

But even more deeply, it's one of these subtle hidden-core-of-rationality things. You know, the sort of thing where I start talking about 'the Way'. It's about *binding yourself to reality.*

In one of Frank Herbert's *Dune* books, IIRC, it is said that a Truthsayer gains their ability to detect lies in others by always speaking truth themselves, so that they form a relationship with the truth whose violation they can feel. It wouldn't work, but I still think it's one of the more beautiful thoughts in fiction. At the very least, to get close to the truth, you have to be willing to press yourself up against reality as tightly as possible, without flinching away, or sneering down.

You can see the bind-yourself-to-reality theme in "Lotteries: A Waste of Hope." Understanding that lottery tickets have negative expected utility, does not mean that you give up the hope of being rich. It means that you stop

wasting that hope on lottery tickets. You put the hope into your job, your school, your startup, your eBay sideline; and if you truly have nothing worth hoping for, then maybe it's time to start looking.

It's not dreams I object to, only *impossible* dreams. The lottery isn't impossible, but it is an un-actionable near-impossibility. It's not that winning the lottery is extremely *difficult* - requires a desperate effort - but that *work*isn't the issue.

I say all this, to exemplify the idea of taking emotional energy that is flowing off to nowhere, and binding it into the realms of reality.

This doesn't mean setting goals that are low enough to be "realistic", i.e., easy and safe and parentally approved. Maybe this is good advice in your personal case, I don't know, but I'm not the one to say it.

What I mean is that you can invest emotional energy in rainbows even if they turn out *not* to be magic. The future is always absurd but it is never *unreal.*

The Hollywood Rationality stereotype is that "rational = emotionless"; the more reasonable you are, the more of your emotions Reason inevitably destroys. In "Feeling Rational" I contrast this against *"That which can be destroyed by the truth should be"* and *"That which the truth nourishes should thrive".* When you have arrived at your best picture of the truth, there is nothing irrational about the emotions you feel as a result of that - the emotions cannot be destroyed by truth, so they must not be irrational.

So instead of *destroying* emotional energies associated with bad explanations for rainbows, as the Hollywood Rationality stereotype would have it, let us *redirect* these emotional energies into reality - bind them to beliefs that are as true as we can make them.

Want to fly? Don't give up on flight. Give up on flying potions and build yourself an airplane.

Remember the theme of "Think Like Reality", where I talked about how when physics seems counterintuitive, you've got to accept that it's not *physics* that's weird, it's *you*?

What I'm talking about now is like that, only with emotions instead of hypotheses - binding your feelings into the real world. Not the "realistic" everyday world. I would be a howling hypocrite if I told you to shut up and do your homework. I mean the *real* real world, the lawful universe, that includes absurdities like Moon landings and the evolution of human intelligence. Just not any magic, anywhere, ever.

It is a Hollywood Rationality meme that "Science takes the fun out of life."

Science puts the fun back *into* life.

Rationality directs your emotional energies into the universe, rather than somewhere else.

If You Demand Magic, Magic Won't Help

> Most witches don't believe in gods. They know that the gods exist, of course. They even deal with them occasionally. But they don't believe in them. They know them too well. It would be like believing in the postman.
> — Terry Pratchett, *Witches Abroad*

Once upon a time, I was pondering the philosophy of fantasy stories

And before anyone chides me for my "failure to understand what fantasy is about", let me say this: I was raised in an SF&F household. I have been reading fantasy stories since I was five years old. I occasionally try to *write* fantasy stories. And I am *not* the sort of person who tries to write for a genre without pondering its philosophy. Where do you think story ideas come from?

Anyway:

I was pondering the philosophy of fantasy stories, and it occurred to me that if there were actually dragons in our world - if you could go down to the zoo, or even to a distant mountain, and meet a fire-breathing dragon - while nobody had ever actually seen a zebra, then our fantasy stories would contain zebras aplenty, while dragons would be unexciting.

Now that's what I call painting yourself into a corner, wot? The grass is always greener on the other side of unreality.

In one of the standard fantasy plots, a protagonist from our Earth, a sympathetic character with lousy grades or a crushing mortgage but still a good heart, suddenly finds themselves in a world where magic operates in place of science. The protagonist often goes on to practice magic, and become in due course a (superpowerful) sorcerer.

Now here's the question - and yes, it is a little unkind, but I think it needs to be asked: Presumably most readers of these novels see themselves in the protagonist's shoes, fantasizing about their own acquisition of sorcery. Wishing for magic. And, barring improbable demographics, most readers of these novels are not scientists.

Born into a world of science, they did not become scientists. What makes them think that, in a world of magic, they would act any differently?

If they don't have the scientific attitude, that nothing is "mere" - the capacity to be interested in merely real things - how will magic help them? If they actually *had* magic, it would be merely *real,* and lose the charm of unattainability. They might be excited at first, but (like the lottery winners who, six months later, aren't nearly as happy as they expected to be), the excitement would soon wear off. Probably as soon as they had to actually *study*spells.

Unless they can find the capacity to take joy in things that are merely real. To be just as excited by hang-gliding, as riding a dragon; to be as excited by making

a light with electricity, as by making a light with magic... even if it takes a little study...

Don't get me wrong. I'm not dissing dragons. Who knows, we might even create some, one of these days.

But if you don't have the capacity to enjoy hang-gliding even though it is *merely real*, then as soon as dragons *turn* real, you're not going to be any more excited by dragons than you are by hang-gliding.

Do you think you would prefer living in the Future, to living in the present? That's a quite understandable preference. Things do seem to be getting better over time.

But don't forget that *this is* the Future, relative to the Dark Ages of a thousand years earlier. You have opportunities undreamt-of even by kings.

If the trend continues, the Future might be a very fine place indeed in which to live. But if you do make it to the Future, what you find, when you get there, will be another Now. If you don't have the basic capacity to enjoy being in a Now - if your emotional energy can *only* go into the Future, if you can *only*hope for a better tomorrow - then no amount of passing time can help you.

(Yes, in the Future there could be a pill that fixes the emotional problem of always looking to the Future. I don't think this invalidates my basic point, which is about what sort of pills we should want to take.)

Matthew C., commenting here on OB, seems very excited about an informally specified "theory" by Rupert Sheldrake which "explains" such non-explanation-demanding phenomena as protein folding and snowflake symmetry. But why isn't Matthew C. just as excited about, say, Special Relativity? Special Relativity is actually *known* to be a law, so why isn't it even *more* exciting? The advantage of becoming excited about a law already known to be true, is that you know your excitement will not be wasted.

If Sheldrake's theory were accepted truth taught in elementary schools, Matthew C. wouldn't care about it. Or why else is Matthew C. fascinated by that one particular law which he believes to be a law of physics, more than all the other laws?

The worst catastrophe you could visit upon the New Age community would be for their rituals to start working reliably, and for UFOs to actually appear in the skies. What would be the point of believing in aliens, if they were just *there,* and everyone else could see them too? In a world where psychic powers were merely real, New Agers wouldn't *believe in* psychic powers, any more than anyone cares enough about gravity to believe in it. (Except for scientists, of course.)

Why am I so negative about magic? Would it be *wrong* for magic to exist?

I'm not actually negative on magic. Remember, I occasionally try to write fantasy stories. But I'm annoyed with this psychology that, if it were born into

a world where spells and potions did work, would pine away for a world where household goods were abundantly produced by assembly lines.

Part of binding yourself to reality, on an emotional as well as intellectual level, is coming to terms with the fact that you *do live here.* Only then can you see this, your world, and whatever opportunities it holds out for you, without wishing your sight away.

Not to put too fine a point on it, but*I've* found no lack of dragons to fight, or magics to master, in this world of my birth. If I were transported into one of those fantasy novels, I wouldn't be surprised to find myself studying the forbidden ultimate sorcery -

- because why should being transported into a magical world change anything? It's not *where* you are, it's *who* you are.

So remember the Litany Against Being Transported Into An Alternate Universe:

If I'm going to be happy anywhere,
Or achieve greatness anywhere,
Or learn true secrets anywhere,
Or save the world anywhere,
Or feel strongly anywhere,
Or help people anywhere,
I may as well do it in reality.

Mundane Magic

As you may recall from some months earlier, I think that part of the rationalist ethos is *binding yourself emotionally* to an absolutely lawful reductionistic universe - a universe containing no ontologically basic mental things such as souls or magic - and pouring all your hope and all your care into that merely real universe and its possibilities, without disappointment.

There's an old trick for combating dukkha where you make a list of things you're grateful for, like a roof over your head.

So why not make a list of abilities you have that would be amazingly cool *if they were magic,* or if only a few chosen individuals had them?

For example, suppose that instead of one eye, you possessed a magical *second* eye embedded in your forehead. And this second eye enabled you to *see into the third dimension* - so that you could somehow tell *how far away* things were - where an ordinary eye would see only a two-dimensional shadow of the true world. Only the possessors of this ability can accurately aim the legendary distance-weapons that kill at ranges far beyond a sword, or use to their fullest potential the shells of ultrafast machinery called "cars".

"Binocular vision" would be too light a term for this ability. We'll only appreciate it once it has a properly impressive name, like Mystic Eyes of Depth Perception.

So here's a list of some of my favorite magical powers:

- *Vibratory Telepathy.* By transmitting invisible vibrations through the very air itself, two users of this ability can *share thoughts.* As a result, Vibratory Telepaths can form emotional bonds much deeper than those possible to other primates.
- *Psychometric Tracery.* By tracing small fine lines on a surface, the Psychometric Tracer can leave impressions of emotions, history, knowledge, even the structure of other spells. This is a higher level than Vibratory Telepathy as a Psychometric Tracer can share the thoughts of long-dead Tracers who lived thousands of years earlier. By reading one Tracery and inscribing another simultaneously, Tracers can duplicate Tracings; and these replicated Tracings can even contain the detailed pattern of other spells and magics. Thus, the Tracers wield almost unimaginable power as magicians; but Tracers can get in trouble trying to use complicated Traceries that they could not have Traced themselves.
- *Multidimensional Kinesis.* With simple, almost unthinking acts of will, the Kinetics can cause extraordinarily complex forces to flow through small tentacles and into any physical object within touching range - not just pushes, but combinations of pushes at many points that can effectively apply torques and twists. The Kinetic ability is far subtler than it first appears: they use it not only to wield existing objects with martial precision, but also to apply forces that sculpt objects into forms more suitable for Kinetic wielding. They even create tools that extend the power of their Kinesis and enable them to sculpt ever-finer and ever-more-complicated tools, a positive feedback loop fully as impressive as it sounds.
- *The Eye.* The user of this ability can perceive infinitesimal traveling twists in the Force that binds matter - tiny vibrations, akin to the life-giving power of the Sun that falls on leaves, but far more subtle. A bearer of the Eye can sense objects far beyond the range of touch using the tiny disturbances they make in the Force. Mountains many days travel away can be known to them as if within arm's reach. According to the bearers of the Eye, when night falls and sunlight fails, they can sense huge fusion fires burning at unthinkable distances - though no one else has any way of verifying this. Possession of a single Eye is said to make the bearer equivalent to royalty.

And finally,

- *The Ultimate Power.* The user of this ability contains a smaller, imperfect echo of the entire universe, enabling them to search out paths through probability to any desired future. If this sounds like a ridiculously powerful ability, you're right - game balance goes right out the window with this one. Extremely rare among life forms, it is the *sekai no ougi* or "hidden technique of the world".

Nothing can oppose the Ultimate Power except the Ultimate Power. Any less-than-ultimate Power will simply be "comprehended" by the Ultimate and disrupted in some inconceivable fashion, or even absorbed into the Ultimates' own power base. For this reason the Ultimate Power is sometimes called the "master technique of techniques" or the "trump card that trumps all other trumps". The more powerful Ultimates can stretch their "comprehension" across galactic distances and aeons of time, and even perceive the bizarre laws of the hidden "world beneath the world".

Ultimates have been killed by immense natural catastrophes, or by extremely swift surprise attacks that give them no chance to use their power. But all such victories are ultimately a matter of luck - it does not confront the Ultimates on their own probability-bending level, and if they survive they will begin to bend Time to avoid future attacks.

But the Ultimate Power itself is also dangerous, and many Ultimates have been destroyed by their own powers - falling into one of the flaws in their imperfect inner echo of the world.

Stripped of weapons and armor and locked in a cell, an Ultimate is still one of the most dangerous life-forms on the planet. A sword can be broken and a limb can be cut off, but the Ultimate Power is "the power that cannot be removed without removing you".

Perhaps because this connection is so intimate, the Ultimates regard one who loses their Ultimate Power permanently - without hope of regaining it - as *schiavo,* or "dead while breathing". The Ultimates argue that the Ultimate Power is so important as to be a necessary part of what makes a creature an end in itself, rather than a means. The Ultimates even insist that anyone who lacks the Ultimate Power cannot begin to truly comprehend the Ultimate Power, and hence, cannot understand why the Ultimate Power is morally important - a suspiciously self-serving argument.

The users of this ability form an absolute aristocracy and treat all other life forms as their pawns.

The Beauty of Settled Science

Facts do not need to be unexplainable, to be beautiful; truths do not become less worth learning, if someone else knows them; beliefs do not become less

worthwhile, if many others share them...

...and if you only care about scientific issues that are controversial, you will end up with a head stuffed full of garbage.

The media thinks that only the cutting edge of science is worth reporting on. How often do you see headlines like "General Relativity still governing planetary orbits" or "Phlogiston theory remains false"? So, by the time anything is solid science, it is no longer a breaking headline. "Newsworthy" science is often based on the thinnest of evidence and wrong half the time - if it were not on the uttermost fringes of the scientific frontier, it would not be breaking news.

Scientific *controversies* are problems *so difficult* that even people who've spent years mastering the field can still fool themselves. That's what makes for the heated arguments that attract all the media attention.

Worse, if you aren't in the field and part of the game, controversies *aren't even fun.*

Oh, sure, you can have the fun of picking a side in an argument. But you can get that in any football game. That's not what the fun of science is about.

Reading a well-written textbook, you get: Carefully phrased explanations for incoming students, math derived step by step (where applicable), plenty of experiments cited as illustration (where applicable), test problems on which to display your new mastery, and a reasonably good guarantee that what you're learning is actually true.

Reading press releases, you usually get: Fake explanations that convey nothing except the delusion of understanding of a result that the press release author didn't understand and that probably has a better-than-even chance of failing to replicate.

Modern science is built on discoveries, built on discoveries, built on discoveries, and so on, all the way back to people like Archimedes, who discovered facts like why boats float, that can make sense even if you don't know about other discoveries. A good place to start traveling that road is at the beginning.

Don't be embarrassed to read *elementary* science textbooks, either. If you want to pretend to be sophisticated, go find a play to sneer at. If you just want to have *fun,* remember that simplicity is at the core of scientific beauty.

And thinking you can jump right into the frontier, when you haven't learned the settled science, is like...

...like trying to climb only the *top* half of Mount Everest (which is the only part that interests you) by standing at the base of the mountain, bending your knees, and jumping *really hard* (so you can pass over the boring parts).

Now I'm not saying that you should never pay attention to scientific controversies. If 40% of oncologists think that white socks cause cancer, and the other 60% violently disagree, this is an important fact to know.

Just don't go thinking that science *has* to be controversial to be interesting.

Or, for that matter, that science has to be recent to be interesting. A steady diet of science *news* is bad for you: You are what you eat, and if you eat only science reporting on fluid situations, without a solid textbook now and then, your brain will turn to liquid.

Amazing Breakthrough Day 1: April 1st

So you're thinking, "April 1st... isn't that already supposed to be April Fool's Day?"

Yes - and that will provide the ideal cover for celebrating Amazing Breakthrough Day.

As I argued in "The Beauty of Settled Science", it is a major problem that media coverage of science focuses only on *breaking news*. Breaking news, in science, occurs at the furthest fringes of the scientific frontier, which means that the new discovery is often:

- Controversial
- Supported by only one experiment
- Way the heck more complicated than an ordinary mortal can handle, and requiring lots of prerequisite science to understand, which is why it wasn't solved three centuries ago
- Later shown to be wrong

People never get to see the *solid* stuff, let alone the *understandable* stuff, because it isn't *breaking news*.

On Amazing Breakthrough Day, I propose, journalists who really care about science can report - under the protective cover of April 1st - such important but neglected science stories as:

- BOATS EXPLAINED: Centuries-Old Problem Solved By Bathtub Nudist
- YOU SHALL NOT CROSS! Knigsberg Tourists' Hopes Dashed
- ARE YOUR LUNGS ON *FIRE*? Link Between Respiration And Combustion Gains Acceptance Among Scientists

Note that every one of these headlines are *true* - they describe events that did, in fact, happen. They just didn't happen *yesterday*.

There have been many humanly understandable amazing breakthroughs in the history of science, which can be understood without a PhD or even BSc. The operative word here is *history.* Think of Archimedes's "Eureka!" when he understood the relation between the water a ship displaces, and the reason the ship floats. This is *far enough back* in scientific history that you don't need to know 50 other discoveries to understand the theory; it can be explained in a couple of graphs; anyone can see how it's useful; and the confirming experiments can be duplicated in your own bathtub.

Modern science is built on discoveries built on discoveries built on discoveries and so on all the way back to Archimedes. Reporting science *only* as breaking news is like wandering into a movie 3/4ths of the way through, writing a story about "Bloody-handed man kisses girl holding gun!" and wandering back out again.

And if your editor says, "Oh, but our readers won't be interested in that -"

Then point out that Reddit and Digg don't link *only* to breaking news. They also link to short webpages that give good explanations of old science. Readers vote it up, and that should tell you something. Explain that if your newspaper doesn't change to look more like Reddit, you'll have to start selling drugs to make payroll. Editors love to hear that sort of thing, right?

On the Internet, a good new explanation of old science *is*news and it spreads like news. Why couldn't the science sections of newspapers work the same way? Why isn't a new *explanation* worth reporting on?

But all this is too visionary for a first step. For now, let's just see if any journalists out there pick up on Amazing Breakthrough Day, where you report on some *understandable* science breakthrough as though it had just occurred.

April 1st. Put it on your calendar.

Is Humanism a Religion-Substitute?

For many years before the Wright Brothers, people dreamed of flying with magic potions. There was nothing irrational about the *raw desire* to fly. There was nothing *tainted* about the wish to look down on a cloud from above. Only the "magic potions" part was irrational.

Suppose you were to put me into an fMRI scanner, and take a movie of my brain's activity levels, while I watched a space shuttle launch. (Wanting to visit space is not "realistic", but it is an essentially lawful dream - one that can be fulfilled in a lawful universe.) The fMRI might - maybe, maybe not - resemble the fMRI of a devout Christian watching a nativity scene.

Should an experimenter obtain this result, there's a lot of people out there, both Christians and some atheists, who would gloat: "Ha, ha, space travel is your religion!"

But that's drawing the wrong category boundary. It's like saying that, because some people once tried to fly by irrational means, no one should ever enjoy looking out of an airplane window on the clouds below.

If a rocket launch is what it takes to give me a feeling of aesthetic transcendence, I do not see this as a *substitute*for religion. That is theomorphism - the viewpoint from gloating religionists who assume that everyone who *isn't* religious has a hole in their mind that wants filling.

Now, to be fair to the religionists, this is not *just*a gloating assumption. There *are* atheists who have religion-shaped holes in their minds. I *have*seen attempts to substitute atheism or even transhumanism for religion. And the result is invariably awful. Utterly awful. Absolutely abjectly awful.

I call such efforts, "hymns to the nonexistence of God".

When someone sets out to write an atheistic hymn - "Hail, oh unintelligent universe," blah, blah, blah - the result will, without exception, suck.

Why? Because they're being imitative. Because they have no motivation for writing the hymn *except* a vague feeling that since churches have hymns, they ought to have one too. And, on a purely artistic level, that puts them far beneath genuine religious art that is not an imitation of anything, but an original expression of emotion.

Religious hymns were (often) written by people who *felt strongly* and*wrote honestly* and put serious effort into the prosody and imagery of their work - that's what gives their work the grace that it possesses, of artistic integrity.

So are atheists doomed to hymnlessness?

There is an acid test of attempts at post-theism. The acid test is: "If religion had never existed among the human species - if we had *never made* the original mistake - would this song, this art, this ritual, this way of thinking, still make sense?"

If humanity had never made the original mistake, there would be no hymns to the nonexistence of God. But there would still be marriages, so the notion of an atheistic marriage ceremony makes perfect sense - as long as you don't suddenly launch into a lecture on how God doesn't exist. Because, in a world where religion *never had* existed, nobody would interrupt a wedding to talk about the implausibility of a distant hypothetical concept. They'd talk about love, children, commitment, honesty, devotion, but who the heck would mention God?

And, in a human world where religion *never had* existed, there would still be people who got tears in their eyes watching a space shuttle launch.

Which is why, even if experiment shows that watching a shuttle launch makes "religion"-associated areas of my brain light up, associated with feelings of transcendence, I do not see that as a *substitute* for religion; I expect the same brain

areas would light up, for the same reason, if I lived in a world where religion had never been invented.

A good "atheistic hymn" is simply a song about anything worth singing about that doesn't happen to be religious.

Also, reversed stupidity is not intelligence. The world's greatest idiot may say the Sun is shining, but that doesn't make it dark out. The point is *not* to create a life that resembles religion as little as possible in every surface aspect - this is the same kind of thinking that inspires hymns to the nonexistence of God. If humanity had never made the original mistake, no one would be *trying to avoid* things that vaguely resembled religion. Believe accurately, then feel accordingly: If space launches actually exist, and watching a rocket rise makes you want to sing, then write the song, dammit.

If I get tears in my eyes at a space shuttle launch, it doesn't mean I'm trying to fill a hole left by religion - it means that my emotional energies, my *caring,* are bound into the real world.

If God did speak plainly, and answer prayers reliably, God would just become one more boringly real thing, no more worth believing in than the postman. If God were real, it would destroy the inner uncertainty that brings forth outward fervor in compensation. And if everyone else believed God were real, it would destroy the specialness of being one of the elect.

If you invest your emotional energy in space travel, you don't have those vulnerabilities. I can *see* the Space Shuttle rise without losing the awe. Everyone else can believe that Space Shuttles are real, and it doesn't make them any less special. I haven't painted myself into the corner.

The choice between God and humanity is not just a choice of drugs. Above all, humanity *actually exists.** *

Scarcity

What follows is taken primarily from Robert Cialdini's *Influence: The Psychology of Persuasion.* I own three copies of this book, one for myself, and two for loaning to friends.

S*carcity,* as that term is used in social psychology, is when things become *more desirable* as they appear *less obtainable.*

- If you put a two-year-old boy in a room with two toys, one toy in the open and the other behind a Plexiglas wall, the two-year-old will ignore the easily accessible toy and go after the apparently forbidden one. If the wall is low enough to be easily climbable, the toddler is no more likely to go after one toy than the other. (Brehm and Weintraub 1977.)

- When Dade County forbade use or possession of phosphate detergents, many Dade residents drove to nearby counties and bought huge amounts of phosphate laundry detergents. Compared to Tampa residents not affected by the regulation, Dade residents rated phosphate detergents as gentler, more effective, more powerful on stains, and even believed that phosphate detergents poured more easily. (Mazis 1975, Mazis et. al. 1973.)

Similarly, information that appears forbidden or secret, seems more important and trustworthy:

- When University of North Carolina students learned that a speech opposing coed dorms had been banned, they became more opposed to coed dorms (without even hearing the speech). (Probably in Ashmore et. al. 1971.)

- When a driver said he had liability insurance, experimental jurors awarded his victim an average of four thousand dollars more than if the driver said he had no insurance. If the judge afterward informed the jurors that information about insurance was inadmissible and must be ignored, jurors awarded an average of thirteen thousand dollars more than if the driver had no insurance. (Broeder 1959.)

- Buyers for supermarkets, told by a supplier that beef was in scarce supply, gave orders for twice as much beef as buyers told it was readily available. Buyers told that beef was in scarce supply, and furthermore, that the information about scarcity was itself scarce - that the shortage was not general knowledge - ordered six times as much beef. (Since the study was conducted in a real-world context, the information provided was in fact correct.) (Knishinsky 1982.)

The conventional theory for explaining this is "psychological reactance", social-psychology-speak for "When you tell people they can't do something, they'll just try even harder." The fundamental instincts involved appear to be preservation of status and preservation of options. We resist dominance, when any human agency tries to restrict our freedom. And when options seem to be in danger of disappearing, even from natural causes, we try to leap on the option before it's gone.

Leaping on disappearing options may be a good adaptation in a hunter-gatherer society - gather the fruits while the tree is still in bloom - but in a money-based society it can be rather costly. Cialdini (1993) reports that in one appliance store he observed, a salesperson who saw that a customer was evincing signs of interest in an appliance would approach, and sadly inform the customer that

the item was out of stock, the last one having been sold only twenty minutes ago. Scarcity creating a sudden jump in desirability, the customer would often ask whether there was any chance that the salesperson could locate an unsold item in the back room, warehouse, or anywhere. "Well," says the salesperson, "that's possible, and I'm willing to check; but do I understand that this is the model you want, and if I can find it at this price, you'll take it?"

As Cialdini remarks, a chief sign of this malfunction is that you dream of *possessing* something, rather than *using* it. (Timothy Ferriss offers similar advice on planning your life: ask which *ongoing experiences* would make you happy, rather than which possessions or status-changes.)

But the really fundamental problem with desiring the unattainable is that as soon as you actually *get* it, it stops being unattainable. If we cannot take joy in the merely available, our lives will *always* be frustrated...

Ashmore, R. D., Ramachandra, V. and Jones, R. A. (1971.) "Censorship as an Attitude Change Induction." Paper presented at Eastern Psychological Association meeting, New York, April 1971.

Brehm, S. S. and Weintraub, M. (1977.) "Physical Barriers and Psychological Reactance: Two-year-olds' Responses to Threats to Freedom." *Journal of Personality and Social Psychology,* **35**: 830–36.

Broeder, D. (1959.) "The University of Chicago Jury Project." *Nebraska Law Review* **38**: 760–74.

Cialdini, R. B. (1993.) *Influence: The Psychology of Persuasion: Revised Edition.* Pp. 237–71. New York: Quill.

Knishinsky, A. (1982.) "The Effects of Scarcity of Material and Exclusivity of Information on Industrial Buyer Perceived Risk in Provoking a Purchase Decision." Doctoral dissertation, Arizona State University.

Mazis, M. B. (1975.) "Antipollution Measures and Psychological Reactance Theory: A Field Experiment." *Journal of Personality and Social Psychology* **31**: 654–66.

Mazis, M. B., Settle, R. B. and Leslie, D. C. (1973.) "Elimination of Phosphate Detergents and Psychological Reactance." *Journal of Marketing Research* **10**: 390–95.

To Spread Science, Keep It Secret

Sometimes I wonder if the Pythagoreans had the right idea.

Yes, I've written about how "science" is inherently public. I've written that "science" is distinguished from merely rational knowledge by the in-principle ability to reproduce scientific experiments for yourself, to know without relying on authority. I've said that "science" should be defined as the publicly accessible knowledge of humankind. I've even suggested that future generations will regard all papers not published in an open-access journal as non-science, i.e., it can't be part of the public knowledge of humankind if you make people pay to read it.

But that's only one vision of the future. In another vision, the knowledge we now call "science" is taken *out* of the public domain - the books and journals hidden away, guarded by mystic cults of gurus wearing robes, requiring fearsome initiation rituals for access - so that more people will *actually* study it.

I mean, right now, people *can* study science but they *don't.*

"Scarcity", it's called in social psychology. What appears to be in limited supply, is more highly valued. And this effect is *especially* strong with information - we're much more likely to try to obtain information that we believe is secret, and to value it more when we do obtain it.

With science, I think, people assume that if the information is freely available, it must not be important. So instead people join cults that have the sense to keep their Great Truths secret. The Great Truth may actually be gibberish, but it's more satisfying than coherent science, because it's *secret.*

Science is the great Purloined Letter of our times, left out in the open and ignored.

Sure, scientific openness helps the scientific elite. They've already *been* through the initiation rituals. But for the rest of the planet, science is kept secret a hundred times more effectively by making it freely available, than if its books were guarded in vaults and you had to walk over hot coals to get access. (This being a fearsome trial indeed, since the great secrets of insulation are only available to Physicist-Initiates of the Third Level.)

If scientific knowledge were hidden in ancient vaults (rather than hidden in inconvenient pay-for-access journals), at least then people would *try* to get into the vaults. They'd be *desperate* to learn science. Especially when they saw the power that Eighth Level Physicists could wield, and were told that they *weren't allowed to know* the explanation.

And if you tried to start a cult around oh, say, Scientology, you'd get some degree of public interest, at first. But people would very quickly start asking uncomfortable questions like "Why haven't you given a public demonstration of your Eighth Level powers, like the Physicists?" and "How come none of the Master Mathematicians seem to want to join your cult?" and "Why should I follow your Founder when he isn't an Eighth Level anything outside his own cult?" and "Why should I study *your* cult *first,* when the Dentists of Doom can do things that are so much more impressive?"

When you look at it from that perspective, the escape of math from the Pythagorean cult starts to look like a major strategic blunder for humanity.

Now, I know what you're going to say: "But science *is* surrounded by fearsome initiation rituals! Plus it's *inherently* difficult to learn! Why doesn't *that* count?" Because the public *thinks* that science is freely available, that's why. If you're *allowed* to learn, it must not be important enough *to* learn.

It's an image problem, people taking their cues from others' attitudes. Just *anyone* can walk into the supermarket and buy a light bulb, and nobody looks at it with awe and reverence. The physics supposedly aren't secret (even though *you* don't know), and there's a one-paragraph explanation in the newspaper that sounds vaguely authoritative and convincing - essentially, no one treats the lightbulb as a sacred mystery, so neither do you.

Even the simplest little things, completely inert objects like crucifixes, can become magical if everyone *looks* at them like they're magic. But since you're theoretically *allowed* to know why the light bulb works without climbing the mountain to find the remote Monastery of Electricians, there's no need to *actually* bother to learn.

Now, because science does in fact have initiation rituals both social and cognitive, scientists are not wholly dissatisfied with their science. The problem is that, in the present world, very few people bother to study science in the first place. Science cannot be the true Secret Knowledge, because just anyone is allowed to know it - *even though, in fact, they don't.*

If the Great Secret of Natural Selection, passed down from Darwin Who Is Not Forgotten, was only ever imparted to you after you paid $2000 and went through a ceremony involving torches and robes and masks and sacrificing an ox, *then*when you were shown the fossils, and shown the optic cable going through the retina under a microscope, and finally told the Truth, you would say "That's the most brilliant thing ever!" and *be satisfied.* After that, if some other cult tried to tell you it was actually a bearded man in the sky 6000 years ago, you'd laugh like hell.

And you know, it might actually be more *fun* to do things that way. Especially if the initiation required you to put together some of the evidence for yourself - together, or with classmates - before you could tell your Science Sensei you were ready to advance to the next level. It wouldn't be *efficient,* sure, but it would be *fun.*

If humanity had never made the mistake - never gone down the religious path, and never learned to fear anything that smacks of religion - then maybe the Ph.D. granting ceremony would involve litanies and chanting, because, hey, that's what people like. Why take the fun out of everything?

Maybe we're just doing it wrong.

And no, I'm not *seriously* proposing that we try to reverse the last five hundred years of openness and classify all the science secret. At least, not at the

moment. Efficiency is important for now, especially in things like medical research. I'm just explaining why it is that I won't tell anyone the Secret of how the ineffable difference between blueness and redness arises from mere atoms for less than $100,000 -

Ahem! I meant to say, I'm telling you about this vision of an alternate Earth, so that you give science equal treatment with cults. So that you don't undervalue scientific truth when you learn it, *just* because it doesn't seem to be protected appropriately to its value. *Imagine* the robes and masks. Visualize yourself creeping into the vaults and stealing the Lost Knowledge of Newton. And don't be fooled by any organization that *does* use robes and masks, unless they also show you the data.

People seem to have holes in their minds for Esoteric Knowledge, Deep Secrets, the Hidden Truth. And I'm not even criticizing this psychology! There *are* deep secret esoteric hidden truths, like quantum mechanics or Bayes-structure. We've just gotten into the habit of presenting the Hidden Truth in a very *unsatisfying* way, wrapped up in false mundanity.

But if the holes for secret knowledge are not filled by true beliefs, they will be filled by false beliefs. There is *nothing but* science to learn - the emotional energy must either be invested in reality, or wasted in total nonsense, or destroyed. For myself, I think it is better to invest the emotional energy; fun should not be needlessly cast away.

Right now, we've got the worst of both worlds. Science isn't *really* free, because the courses are expensive and the textbooks are expensive. But the public *thinks* that anyone is allowed to know, so it must not be important.

Ideally, you would want to arrange things the other way around.

Initiation Ceremony

The torches that lit the narrow stairwell burned intensely and in the wrong color, flame like melting gold or shattered suns.
192... 193...*
Brennan's sandals clicked softly on the stone steps, snicking in sequence, like dominos very slowly falling.
227... 228...*
Half a circle ahead of him, a trailing fringe of dark cloth whispered down the stairs, the robed figure itself staying just out of sight.
239... 240...
Not much longer, Brennan predicted to himself, and his guess was accurate:
Sixteen times sixteen steps was the number, and they stood before the portal of glass.
The great curved gate had been wrought with cunning, humor, and close attention to indices of refraction: it warped light, bent it, folded it, and generally

abused it, so that there were hints of what was on the other side (stronger light sources, dark walls) but no possible way of *seeing through* - unless, of course, you had the key: the counter-door, thick for thin and thin for thick, in which case the two would cancel out.
From the robed figure beside Brennan, two hands emerged, gloved in reflective cloth to conceal skin's color. Fingers like slim mirrors grasped the handles of the warped gate - handles that Brennan had not guessed; in all that distortion, shapes could only be anticipated, not seen.
"Do you want to know?" whispered the guide; a whisper nearly as loud as an ordinary voice, but not revealing the slightest hint of gender.
Brennan paused. The answer to the question seemed suspiciously, indeed extraordinarily obvious, even for ritual.

"Yes," Brennan said finally.
The guide only regarded him silently.
"Yes, I want to know," said Brennan.
"Know *what,* exactly?" whispered the figure.
Brennan's face scrunched up in concentration, trying to visualize the game to its end, and hoping he hadn't blown it already; until finally he fell back on the first and last resort, which is the truth:
"It doesn't matter," said Brennan, "the answer is still yes."
The glass gate parted down the middle, and slid, with only the tiniest scraping sound, into the surrounding stone.
The revealed room was lined, wall-to-wall, with figures robed and hooded in light-absorbing cloth. The straight walls were not themselves black stone, but mirrored, tiling a square grid of dark robes out to infinity in all directions; so that it seemed as if the people of some much vaster city, or perhaps the whole human kind, watched in assembly. There was a hint of moist warmth in the air of the room, the breath of the gathered: a scent of crowds.
Brennan's guide moved to the center of the square, where burned four torches of that relentless yellow flame. Brennan followed, and when he stopped, he realized with a slight shock that all the cowled hoods were now looking directly at him. Brennan had never before in his life been the focus of such absolute attention; it was frightening, but not entirely unpleasant.
"He is here," said the guide in that strange loud whisper.
The endless grid of robed figures replied in one voice: perfectly blended, exactly synchronized, so that not a single individual could be singled out from the rest, and betrayed:
"*Who is absent?*"
"Jakob Bernoulli," intoned the guide, and the walls replied:
"*Is dead but not forgotten.*"
"Abraham de Moivre,"
"*Is dead but not forgotten.*"
"Pierre-Simon Laplace,"
"*Is dead but not forgotten.*"
"Edwin Thompson Jaynes,"

"*Is dead but not forgotten.*"
"They died," said the guide, "and they are lost to us; but we still have each other, and the project continues."
In the silence, the guide turned to Brennan, and stretched forth a hand, on which rested a small ring of nearly transparent material.
* *Brennan stepped forward to take the ring -*
But the hand clenched tightly shut.
"If three-fourths of the humans in this room are women," said the guide, "and three-fourths of the women and half of the men belong to the Heresy of Virtue, and I am a Virtuist, what is the probability that I am a man?"
"Two-elevenths," Brennan said confidently.
There was a moment of absolute silence.
Then a titter of shocked laughter.
The guide's whisper came again, truly quiet this time, almost nonexistent: "It's one-sixth, actually."
Brennan's cheeks were flaming so hard that he thought his face might melt off. The instinct was very strong to run out of the room and up the stairs and flee the city and change his name and start his life over again and get it right this time.
"An honest mistake is at least honest," said the guide, louder now, "and we may know the honesty by its relinquishment. If I am a Virtuist, what is the probability that I am a man?"
"One -" Brennan started to say.
Then he stopped. Again, the horrible silence.
"Just say 'one-sixth' already," stage-whispered the figure, this time loud enough for the walls to hear; then there was more laughter, not all of it kind.
*Brennan was breathing rapidly and there was sweat on his forehead. If he was wrong about this, he really*was* going to flee the city. "Three fourths women times three fourths Virtuists is nine sixteenths female Virtuists in this room. One fourth men times one half Virtuists is two sixteenths male Virtuists. If I have only that information and the fact that you are a Virtuist, I would then estimate odds of two to nine, or a probability of two-elevenths, that you are male. Though I do not, in fact, believe the information given is correct. For one thing, it seems too neat. For another, there are an odd number of people in this room."
The hand stretched out again, and opened.
Brennan took the ring. It looked almost invisible, in the torchlight; not glass, but some material with a refractive index very close to air. The ring was warm from the guide's hand, and felt like a tiny living thing as it embraced his finger. The relief was so great that he nearly didn't hear the cowled figures applauding. From the robed guide came one last whisper:
"You are now a novice of the Bayesian Conspiracy."

Awww, a Zebra

This image recently showed up on Flickr (original is nicer):

With the caption:

> *"Alas for those who turn their eyes from zebras and dream of dragons! If we cannot learn to take joy in the merely real, our lives shall be empty indeed." — Eliezer S. Yudkowsky.*

"Awww!", I said, and called over my girlfriend over to look.

"Awww!", she said, and then looked at me, and said, "I think you need to take your own advice!"

Me: "But I'm looking at the zebra!"
Her: "*On a computer!*"
Me: *(Turns away, hides face.)*
Her: "Have you ever even *seen* a zebra in real life?"
Me: "Yes! Yes, I have! My parents took me to Lincoln Park Zoo! ... man, I hated that place."

Hand vs Fingers

Back to our original topic: Reductionism, which (in case you've forgotten) is part of a sequence on the Mind Projection Fallacy. There can be emotional problems in accepting reductionism, if you think that things have to be fundamental to be fun. But this position commits us to never taking joy in anything more complicated than a quark, and so I prefer to reject it.

To review, the reductionist thesis is that we use multi-level models for computational reasons, but physical reality has only a single level. If this doesn't sound familiar, please reread "Reductionism".

Today I'd like to pose the following conundrum: When you pick up a cup of water, is it your *hand* that picks it up?

Most people, of course, go with the naive popular answer: "Yes."

Recently, however, scientists have made a stunning discovery: It's not your *hand* that holds the cup, it's actually your fingers, thumb, and palm.

Yes, I know! I was shocked too. But it seems that after scientists measured the forces exerted on the cup by each of your fingers, your thumb, and your palm,

they found there was no force left over - so the force exerted by your *hand* must be zero.

The theme here is that, if you can *see how* (not just *know that*) a higher level reduces to a lower one, they will not seem like separate things within your map; you will be able to *see* how silly it is to think that your fingers could be in one place, and your hand somewhere else; you will be able to *see* how silly it is to argue about whether it is your hand picks up the cup, or your fingers.

The operative word is "see", as in concrete visualization. Imagining your hand causes you to imagine the fingers and thumb and palm; conversely, imagining fingers and thumb and palm causes you to identify a hand in the mental picture. Thus the high level *of your map* and the low level *of your map* will be tightly bound together *in your mind*.

In reality, of course, the levels are bound together even tighter than that - bound together by the tightest possible binding: physical identity. You can *see* this: You can *see* that saying (1) "hand" or (2) "fingers and thumb and palm", does not refer to different *things,* but different *points of view*.

But suppose you lack the knowledge to so tightly bind together the levels of your map. For example, you could have a "hand scanner" that showed a "hand" as a dot on a map (like an old-fashioned radar display), and similar scanners for fingers/thumbs/palms; then you would see a cluster of dots around the hand, but you would be able to *imagine* the hand-dot moving off from the others. So, even though the physical reality of the hand (that is, the thing the dot corresponds to) was identical with / strictly composed of the physical realities of the fingers and thumb and palm, you would not be able to see this fact; even if someone told you, or you guessed from the correspondence of the dots, you would only *know* the fact of reduction, not *see* it. You would still be able to *imagine* the hand dot moving around independently, even though, if the physical makeup of the sensors were held constant, it would be physically impossible for this to actually happen.

Or, at a still lower level of binding, people might just tell you "There's a hand over there, and some fingers over there" - in which case you would know little more than a Good-Old-Fashioned AI representing the situation using suggestively named LISP tokens. There wouldn't be anything *obviously* contradictory about asserting:

```
|- Inside(Room,Hand)<br>|- ~Inside(Room,Fingers)
```

because you would not possess the *knowledge*

```
|- Inside(x, Hand) -> Inside(x,Fingers)
```

None of this says that a hand can actually detach its existence from your fingers and crawl, ghostlike, across the room; it just says that a Good-Old-Fashioned

AI with a propositional representation may not *know* any better. The map is not the territory.

In particular, you shouldn't draw too many conclusions from how it seems *conceptually possible,* in the mind of some specific conceiver, to separate the hand from its constituent elements of fingers, thumb, and palm. Conceptual possibility is not the same as logical possibility or physical possibility.

It is *conceptually* possible *to you* that 235757 is prime, because you don't know any better. But it isn't *logically* possible that 235757 is prime; if you were logically omniscient, 235757 would be obviously composite (and you would know the factors). That that's why we have the notion of impossible possible worlds, so that we can put probability distributions on propositions that may or may not be *in fact* logically impossible.

And you can imagine philosophers who criticize "eliminative fingerists" who contradict the direct facts of experience - we can *feel* our hand holding the cup, after all - by suggesting that "hands" *don't really exist,* in which case, obviously, the cup would fall down. And philosophers who suggest "appendigital bridging laws" to explain how a particular configuration of fingers, evokes a hand into existence - with the note, of course, that while our world contains those particular appendigital bridging laws, the laws could have been conceivably different, and so are not in any sense *necessary facts,* etc.

All of these are cases of Mind Projection Fallacy, and what I call "naive philosophical realism" - the confusion of philosophical intuitions for direct, veridical information about reality. Your inability to imagine something is just a computational fact about what your brain can or can't imagine. Another brain might work differently.

Angry Atoms

Fundamental physics - quarks 'n stuff - is far removed from the levels we can *see,* like hands and fingers. At best, you can know how to replicate the experiments which show that your hand (like everything else) is composed of quarks, and you may know how to derive a few equations for things like atoms and electron clouds and molecules.

At worst, the existence of quarks beneath your hand may just be something you were told. In which case it's questionable in one what sense you can be said to "know" it at all, even if you repeat back the same word "quark" that a physicist would use to convey knowledge to another physicist.

Either way, you can't actually *see* the identity between levels - no one has a brain large enough to *visualize* avogadros of quarks and recognize a hand-pattern in them.

But we at least understand what hands *do.* Hands push on things, exert forces on them. When we're told about atoms, we visualize little billiard balls bumping into each other. This makes it seem obvious that "atoms" can push on things too, by bumping into them.

Now this notion of atoms is not quite correct. But so far as *human imagination* goes, it's relatively easy to imagine our hand being made up of a little galaxy of swirling billiard balls, pushing on things when our "fingers" touch them. Democritus imagined this 2400 years ago, and there was a time, roughly 1803–1922, when Science thought he was right.

But what about, say, anger?

How could little billiard balls be angry? Tiny frowny faces on the billiard balls?

Put yourself in the shoes of, say, a hunter-gatherer - someone who may not even have a notion of writing, let alone the notion of using base matter to perform computations - someone who has no idea that such a thing as neurons exist. Then you can imagine the *functional* gap that your ancestors might have perceived between billiard balls and "Grrr! Aaarg!"

Forget about subjective experience for the moment, and consider the sheer *behavioral* gap between anger and billiard balls. The difference between what little billiard balls *do,* and what anger makes people *do.* Anger can make people raise their fists and hit someone - or say snide things behind their backs - or plant scorpions in their tents at night. Billiard balls just push on things.

Try to put yourself in the shoes of the hunter-gatherer who's never had the "Aha!" of information-processing. Try to avoid hindsight bias about things like neurons and computers. Only then will you be able to see the uncrossable explanatory gap:

How can you explain angry behavior in terms of billiard balls?

Well, the *obvious* materialist conjecture is that the little billiard balls push on your arm and make you hit someone, or push on your tongue so that insults come out.

But how do the little billiard balls know how to do this - or how to guide your tongue and fingers through long-term plots - if they aren't angry themselves?

And besides, if you're not seduced by - gasp! - scientism, you can see from a first-person perspective that this explanation is obviously false. Atoms can push on your arm, but they can't make you *want* anything.

Someone may point out that drinking wine can make you angry. But who says that wine is made exclusively of little billiard balls? Maybe wine just contains a potency of angerness.

Clearly, reductionism is just a flawed notion.

(The novice goes astray and says "The art failed me"; the master goes astray and says "I failed my art.")

What does it take to cross this gap? It's not just the idea of "neurons" that "process information" - if you say only this and nothing more, it just inserts a magical, unexplained level-crossing rule into your model, where you go from billiards to thoughts.

But an Artificial Intelligence programmer who knows how to create a chess-playing program out of base matter, has taken a *genuine* step toward crossing the gap. If you understand concepts like consequentialism, backward chaining, utility functions, and search trees, you can make merely causal/mechanical systems compute plans.

The trick goes something like this: For each possible chess move, compute the moves your opponent could make, then your responses to those moves, and so on; evaluate the furthest position you can see using some local algorithm (you might simply count up the material); then trace back using minimax to find the best move on the current board; then make that move.

More generally: If you have chains of causality inside the mind that have a kind of mapping - a mirror, an echo - to what goes on in the environment, then you can run a utility function over the end products of imagination, and find an action that achieves something which the utility function rates highly, and output that action. It is not necessary for the chains of causality inside the mind, that are similar to the environment, to be made out of billiard balls that have little auras of intentionality. Deep Blue's transistors do not need little chess pieces carved on them, in order to work. See also The Simple Truth.

All this is still tremendously oversimplified, but it should, at least, reduce the apparent length of the gap. If you can understand all that, you can see how a planner built out of base matter can be influenced by alcohol to output more angry behaviors. The billiard balls in the alcohol push on the billiard balls making up the utility function.

But even if you know how to write small AIs, you can't *visualize* the level-crossing between transistors and chess. There are too many transistors, and too many moves to check.

Likewise, even if you knew all the facts of neurology, you would not be able to *visualize* the level-crossing between neurons and anger - let alone the level-crossing between atoms and anger. Not the way you can visualize a hand consisting of fingers, thumb, and palm.

And suppose a cognitive scientist just flatly tells you "Anger is hormones"? Even if you repeat back the words, it doesn't mean you've crossed the gap. You may believe you believe it, but that's not the same as understanding what little billiard balls have to do with wanting to hit someone.

So you come up with interpretations like, "Anger is *mere* hormones, it's caused by little molecules, so it must not be justified in any moral sense - *that's* why you should learn to control your anger."

Or, "There isn't really any such thing as anger - it's an illusion, a quotation with no referent, like a mirage of water in the desert, or looking in the garage for a dragon and not finding one."

These are both tough pills to swallow (not that you *should* swallow them) and so it is a good easier to profess them than to believe them.

I think this is what non-reductionists/non-materialists think they are criticizing when they criticize reductive materialism.

But materialism isn't that easy. It's not as cheap as saying, "Anger is made out of atoms - there, now I'm done." That wouldn't explain how to get from billiard balls to hitting. You need the specific insights of computation, consequentialism, and search trees before you can start to close the explanatory gap.

All this was a relatively easy example *by modern standards,* because I restricted myself to talking about angry *behaviors.* Talking about outputs doesn't require you to appreciate how an algorithm feels from inside (cross a first-person/third-person gap) or dissolve a wrong question (untangle places where the interior of your own mind runs skew to reality).

Going from material substances that bend and break, burn and fall, push and shove, to angry *behavior,* is just a practice problem by the standards of modern philosophy. But it is an *important* practice problem. It can only be fully appreciated, if you realize how *hard* it would have been to solve before writing was invented. There was once an explanatory gap here - though it may not seem that way in hindsight, now that it's been bridged for generations.

Explanatory gaps can be crossed, if you accept help from science, and don't trust the view from the interior of your own mind.

Heat vs Motion

After yesterday's post, it occurred to me that there's a much simpler example of reductionism jumping a gap of apparent-difference-in-kind: the reduction of heat to motion.

Today, the equivalence of heat and motion may seem too obvious in hindsight - everyone says that "heat is motion", therefore, it can't be a "weird" belief.

But there was a time when the kinetic theory of heat was a highly controversial scientific hypothesis, contrasting to belief in a caloric fluid that flowed from hot objects to cold objects. Still earlier, the main theory of heat was "Phlogiston!"

Suppose you'd *separately* studied kinetic theory and caloric theory. You now know something about kinetics: collisions, elastic rebounds, momentum, kinetic energy, gravity, inertia, free trajectories. Separately, you know something

about heat: Temperatures, pressures, combustion, heat flows, engines, melting, vaporization.

Not only is this state of knowledge a plausible one, it is the state of knowledge possessed by e.g. Sadi Carnot, who, working strictly from within the caloric theory of heat, developed the principle of the Carnot cycle - a heat engine of maximum efficiency, whose existence implies the second law of thermodynamics. This in 1824, when kinetics was a highly developed science.

Suppose, like Carnot, you know a great deal about kinetics, and a great deal about heat, as *separate* entities. Separate entities *of knowledge*, that is: your brain has separate filing baskets for beliefs about kinetics and beliefs about heat. But from the inside, this state of knowledge *feels* like living in a world of moving things and hot things, a world where motion and heat are independent properties of matter.

Now a Physicist From The Future comes along and tells you: "Where there is heat, there is motion, and vice versa. That's why, for example, rubbing things together makes them hotter."

There are (at least) two possible interpretations you could attach to this statement, "Where there is heat, there is motion, and vice versa."

First, you could suppose that heat and motion exist separately - that the caloric theory is correct - but that among our universe's physical laws is a "bridging law" which states that, where objects are moving quickly, caloric will come into existence. And conversely, another bridging law says that caloric can exert pressure on things and make them move, which is why a hotter gas exerts more pressure on its enclosure (thus a steam engine can use steam to drive a piston).

Second, you could suppose that heat and motion are, in some as-yet-mysterious sense,*the same thing*.

"Nonsense," says Thinker 1, "the words 'heat' and 'motion' have two different meanings; that is why we have two different words. We know how to determine when we will call an observed phenomenon 'heat' - heat can melt things, or make them burst into flame. We know how to determine when we will say that an object is 'moving quickly' - it changes position; and when it crashes, it may deform, or shatter. Heat is concerned with change of substance; motion, with change of position and shape. To say that these two words have the same meaning is simply to confuse yourself."

"Impossible," says Thinker 2. "It may be that, in our world, heat and motion are associated by bridging laws, so that it is a law of physics that motion creates caloric, and vice versa. But I can easily imagine a world where rubbing things together does *not* make them hotter, and gases *don't* exert more pressure at higher temperatures. Since there are possible worlds where heat and motion are not associated, they must be different properties - this is true a priori."

Thinker 1 is confusing the quotation and the referent. 2 + 2 = 4, but "2 + 2" != "4". The string "2 + 2" contains 5 characters (including whitespace)

and the string "4" contains only 1 character. If you type the two strings into a Python interpreter, they yield the same output, -> 4. So you can't conclude, from looking at the strings "2 + 2" and "4", that just because the strings are different, they must have different "meanings" relative to the Python Interpreter.

The words "heat" and "kinetic energy" can be said to "refer to" the same thing, even before we *know* how heat reduces to motion, in the sense that we don't know yet what the reference is, but the references are in fact the same. You might imagine an Idealized Omniscient Science Interpreter that would give the same output when we typed in "heat" and "kinetic energy" on the command line.

I talk about the Science Interpreter to emphasize that, to dereference the pointer, you've got to step outside cognition. The end result of the dereference is something out there in reality, not in anyone's mind. So you can *say* "real referent" or "actual referent", but you can't *evaluate* the words locally, from the inside of your own head. You can't reason using the actual heat-referent - if you thought using *real heat*, thinking "1 million Kelvin" would vaporize your brain. But, by forming a belief about your belief about heat, you can talk *about* your belief about heat, and say things like "It's possible that my belief about heat doesn't much resemble *real* heat." You can't actually perform that comparison right there in your own mind, but you can talk *about* it.

Hence you can say, "My beliefs about heat and motion are not the same beliefs, but it's possible that actual heat and actual motion are the same thing." It's just like being able to acknowledge that "the morning star" and "the evening star" might be the same planet, while also understanding that you can't determine this just by examining your beliefs - you've got to haul out the telescope.

Thinker 2's mistake follows similarly. A physicist told him, "Where there is heat, there is motion" and P2 mistook this for a statement of *physical law*: The presence of caloric *causes* the existence of motion. What the physicist really means is more akin to an *inferential rule:* Where you are told there is "heat", deduce the presence of "motion".

From this basic projection of a multilevel model into a multilevel reality follows another, distinct error: the conflation of conceptual possibility with logical possibility. To Sadi Carnot, it is *conceivable* that there could be another world where heat and motion are not associated. To Richard Feynman, armed with specific knowledge of how to derive equations about heat from equations about motion, this idea is not only inconceivable, but so wildly inconsistent as to make one's head explode.

I should note, in fairness to philosophers, that there are philosophers who have said these things. For example, Hilary Putnam, writing on the "Twin Earth" thought experiment:

> Once we have discovered that water (in the actual world) is H20,

> *nothing counts as a possible world in which water isn't H20.* In particular, if a "logically possible" statement is one that holds in some "logically possible world", *it isn't logically possible that water isn't H20.*
>
> On the other hand, we can perfectly well imagine having experiences that would convince us (and that would make it rational to believe that) water *isn't* H20. In that sense, it is conceivable that water isn't H20. It is conceivable but it isn't logically possible! Conceivability is no proof of logical possibility.

It appears to me that "water" is being used in two different senses in these two paragraphs - one in which the word "water" *refers* to what we type into the Science Interpreter, and one in which "water" *refers* to what we get out of the Science Interpreter when we type "water" into it. In the first paragraph, Hilary seems to be saying that after we do some experiments and find out that water is H20, water becomes automatically redefined to *mean* H20. But you could coherently hold a different position about whether the word "water" now *means* "H20" or "whatever is *really* in that bottle next to me", so long as you use your terms consistently.

I believe the above has already been said as well? Anyway...

It is quite possible for there to be only *one* thing out-there-in-the-world, but for it to take on sufficiently different forms, and for you yourself to be sufficiently ignorant of the reduction, that it feels like living in a world containing two entirely different things. Knowledge concerning these two different phenomena may taught in two different classes, and studied by two different academic fields, located in two different buildings of your university.

You've got to put yourself quite a ways back, into a historically realistic frame of mind, to remember how *different* heat and motion once seemed. Though, depending on how much you know today, it may not be as hard as all that, if you can look past the pressure of conventionality (that is, "heat is motion" is an un-weird belief, "heat is not motion" is a weird belief). I mean, suppose that tomorrow the physicists stepped forward and said, "Our popularizations of science have always contained one lie. Actually, heat has nothing to do with motion." Could you *prove* they were wrong?

Saying "Maybe heat and motion are the same thing!" is easy. The difficult part is explaining *how*. It takes a great deal of detailed knowledge to get yourself to the point where you can no longer *conceive* of a world in which the two phenomena go separate ways. Reduction isn't cheap, and that's why it buys so much.

Or maybe you could say: "Reductionism is easy, reduction is hard." But it does kinda help to be a reductionist, I think, when it comes time to go looking for a reduction.

Brain Breakthrough! It's Made of Neurons!

amazing breakthrough, a multinational team of scientists led by Nobel laureate Santiago Ramon y Cajal announced that the brain is composed of a *ridiculously* complicated network of tiny cells connected to each other by infinitesimal threads and branches.

The multinational team - which also includes the famous technician Antonie van Leeuwenhoek, and possibly Imhotep, promoted to the Egyptian god of medicine - issued this statement:

"The present discovery culminates years of research indicating that the convoluted squishy thing inside our skulls is even more complicated than it looks. Thanks to Cajal's application of a new staining technique invented by Camillo Golgi, we have learned that this structure is not a continuous network like the blood vessels of the body, but is actually composed of many tiny cells, or"neurons", connected to one another by even more tiny filaments.

"Other extensive evidence, beginning from Greek medical researcher Alcmaeon and continuing through Paul Broca's research on speech deficits, indicates that the brain is the seat of reason.

"Nemesius, the Bishop of Emesia, has previously argued that brain tissue is too earthy to act as an intermediary between the body and soul, and so the mental faculties are located in the ventricles of the brain. However, if this is correct, there is no reason why this organ should turn out to have an immensely complicated internal composition.

"Charles Babbage has independently suggested that many small mechanical devices could be collected into an 'Analytical Engine', capable of performing activities, such as arithmetic, which are widely believed to require thought. The work of Luigi Galvani and Hermann von Helmholtz suggests that the activities of neurons are electrochemical in nature, rather than mechanical pressures as previously believed. Nonetheless, we think an analogy with Babbage's 'Analytical Engine' suggests that a vastly complicated network of neurons could similarly exhibit thoughtful properties.

"We have found an enormously complicated material system located where the mind should be. The implications are shocking, and must be squarely faced. We believe that the present research offers strong experimental evidence that Benedictus Spinoza was correct, and Rene Descartes wrong: Mind and body are of one substance.

"In combination with the work of Charles Darwin showing how such a complicated organ could, in principle, have arisen as the result of processes not themselves intelligent, the bulk of scientific evidence now seems to indicate that intelligence is ontologically non-fundamental and has an extended origin in time. This strongly weighs against theories which assign mental entities an

ontologically fundamental or causally primal status, including all religions ever invented.

"Much work remains to be done on discovering the specific identities between electrochemical interactions between neurons, and thoughts. Nonetheless, we believe our discovery offers the promise, though not yet the realization, of a full scientific account of thought. The problem may now be declared, if not solved, then solvable."

We regret that Cajal and most of the other researchers involved on the Project are no longer available for comment.

Reductive Reference

The reductionist thesis (as I formulate it) is that human minds, for reasons of efficiency, use a multi-level map in which we separately *think* about things like "atoms" and "quarks", "hands" and "fingers", or "heat" and "kinetic energy". Reality itself, on the other hand, is single-level in the sense that it does not seem to contain atoms as *separate, additional, causally efficacious* entities *over and above* quarks.

Sadi Carnot formulated the (precursor to) the second law of thermodynamics using the caloric theory of heat, in which heat was just a fluid that flowed from hot things to cold things, produced by fire, making gases expand - the effects of heat were studied separately from the science of kinetics, considerably before the reduction took place. If you're trying to design a steam engine, the effects of all those tiny vibrations and collisions which we name "heat" can be summarized into a much simpler description than the full quantum mechanics of the quarks. Humans compute efficiently, thinking of only significant effects on goal-relevant quantities.

But reality itself does seem to use the full quantum mechanics of the quarks. I once met a fellow who thought that if you used General Relativity to compute a low-velocity problem, like an artillery shell, GR would give you the *wrong answer* - not just a slow answer, but an *experimentally wrong* answer - because at low velocities, artillery shells are governed by Newtonian mechanics, not GR. This is exactly how physics does *not* work. Reality just seems to go on crunching through General Relativity, even when it only makes a difference at the fourteenth decimal place, which a human would regard as a huge waste of computing power. Physics does it with brute force. No one has *ever*caught physics simplifying its calculations - or if someone did catch it, the Matrix Lords erased the memory afterward.

Our map, then, is very much unlike the territory; our maps are multi-level, the territory is single-level. Since the representation is so incredibly unlike the referent, in what sense can a belief like "I am wearing socks" be called *true,* when in reality itself, there are only quarks?

In case you've forgotten what the word "true" means, the classic definition was given by Alfred Tarski:

> The statement "snow is white" is *true* if and only if snow is white.

In case you've forgotten what the difference is between the statement "I believe 'snow is white'" and "'Snow is white' is true", see here. Truth can't be evaluated *just* by looking inside your own head - if you want to know, for example, whether "the morning star = the evening star", you need a telescope; it's not enough just to look at the beliefs themselves.

This is the point missed by the postmodernist folks screaming, "But how do you *know* your beliefs are true?" When you do an experiment, you actually *are* going outside your own head. You're engaging in a complex interaction whose outcome is causally determined by the thing you're reasoning about, not just your beliefs about it. I once defined "reality" as follows:

> Even when I have a simple hypothesis, strongly supported by all the evidence I know, sometimes I'm still surprised. So I need different names for the thingies that determine my predictions and the thingy that determines my experimental results. I call the former thingies 'belief', and the latter thingy 'reality'."

The interpretation of your experiment still depends on your prior beliefs. I'm not going to talk, for the moment, about Where Priors Come From, because that is not the subject of this blog post. My point is that truth refers to an *ideal* comparison between a belief and reality. Because we understand that planets are distinct from beliefs about planets, we can design an experiment to test whether the belief "the morning star and the evening star are the same planet" is *true.* This experiment will involve telescopes, not just introspection, because we understand that "truth" involves comparing an internal belief to an external fact; so we use an instrument, the telescope, whose perceived behavior we believe to depend on the external fact of the planet.

Believing that the telescope helps us evaluate the "truth" of "morning star = evening star", relies on our prior beliefs about the telescope interacting with the planet. Again, I'm not going to address that in this particular blog post, except to quote one of my favorite Raymond Smullyan lines: "If the more sophisticated reader objects to this statement on the grounds of its being a mere tautology, then please at least give the statement credit for not being inconsistent." Similarly, I don't see the use of a telescope as circular logic, but as reflective coherence; for every systematic way of arriving at truth, there ought to be a rational explanation for how it works.

The question on the table is what it *means* for "snow is white" to be *true,* when, in reality, there are just quarks.

There's a certain pattern of neural connections making up your beliefs about "snow" and "whiteness" - we believe this, but we do not know, and cannot concretely visualize, the actual neural connections. Which are, themselves, embodied in a pattern of quarks even less known. Out there in the world, there are water molecules whose temperature is low enough that they have arranged themselves in tiled repeating patterns; they look nothing like the tangles of neurons. In what sense, comparing one (ever-fluctuating) pattern of quarks to the other, is the belief "snow is white" *true?*

Obviously, neither I nor anyone else can offer an Ideal Quark Comparer Function that accepts a quark-level description of a neurally embodied belief (including the surrounding brain) and a quark-level description of a snowflake (and the surrounding laws of optics), and outputs "true" or "false" over "snow is white". And who says the fundamental level is *really* about particle fields?

On the other hand, throwing out all beliefs because they aren't written as gigantic unmanageable specifications about quarks we can't even see... doesn't seem like a very prudent idea. Not the best way to optimize our goals.

It seems to me that a word like "snow" or "white" can be taken as a kind of promissory note - not a *known* specification of exactly which physical quark configurations count as "snow", but, nonetheless, there are things you call snow and things you don't call snow, and even if you got a few items wrong (like plastic snow), an Ideal Omniscient Science Interpreter would see a tight cluster in the center and redraw the boundary to have a simpler definition.

In a single-layer universe whose bottom layer is unknown, or uncertain, or just too large to talk about, the concepts in a multi-layer mind can be said to represent a kind of promissory note - we don't know *what* they correspond to, out there. But it seems to us that we can distinguish positive from negative cases, in a predictively productive way, so we think - perhaps in a fully general sense - that there is *some* difference of quarks, *some* difference of configurations at the fundamental level, which explains the differences that feed into our senses, and ultimately result in our saying "snow" or "not snow".

I see this white stuff, and it is the same on several occasions, so I hypothesize a stable latent cause in the environment - I give it the name "snow"; "snow" is then a promissory note referring to a believed-in simple boundary that could be drawn around the unseen causes of my experience.

Hilary Putnam's "Twin Earth" thought experiment, where water is not H20 but some strange other substance denoted XYZ, otherwise behaving much like water, and the subsequent philosophical debate, helps to highlight this issue. "Snow" doesn't have a logical definition known to us - it's more like an empirically determined pointer to a logical definition. This is true even if you believe that snow is ice crystals is low-temperature tiled water molecules. The water molecules are made of quarks. What if quarks turn out to be made of something else? What *is* a snowflake, then? You don't know - but it's still a snowflake, not a fire hydrant.

And of course, these very paragraphs I have just written, are likewise far above the level of quarks. "Sensing white stuff, visually categorizing it, and thinking 'snow' or 'not snow'" - this is also talking very far above the quarks. So my meta-beliefs are also promissory notes, for things that an Ideal Omniscient Science Interpreter might know about which configurations of the quarks (or whatever) making up my brain, correspond to "believing 'snow is white'".

But then, the entire grasp that we have upon reality, is made up of promissory notes of this kind. So, rather than calling it circular, I prefer to call it self-consistent.

This can be a bit unnerving - maintaining a precarious epistemic perch, in both object-level beliefs and reflection, far above a huge unknown underlying fundamental reality, and hoping one doesn't fall off.

On reflection, though, it's hard to see how things could be any other way.

So at the end of the day, the statement "reality does not contain hands as fundamental, additional, separate causal entities, over and above quarks" is not the same statement as "hands do not exist" or "I don't have any hands". There are no *fundamental* hands; hands are made of fingers, palm, and thumb, which in turn are made of muscle and bone, all the way down to elementary particle fields, which are the fundamental causal entities, so far as we currently know.

This is not the same as saying, "there are no 'hands'." It is not the same as saying, "the word 'hands' is a promissory note that will never be paid, because there is no empirical cluster that corresponds to it"; or "the 'hands' note will never be paid, because it is logically impossible to reconcile its supposed characteristics"; or "the statement 'humans have hands' refers to a sensible state of affairs, but reality is not in that state".

Just: There are patterns that exist *in* reality where we see "hands", and these patterns have something in common, but they are not fundamental.

If I *really* had no hands - if reality suddenly transitioned to be in a state that we would describe as "Eliezer has no hands" - reality would shortly thereafter correspond to a state we would describe as "Eliezer screams as blood jets out of his wrist stumps".

And this is *true,* even though the above paragraph hasn't specified any quark positions.

The previous sentence is likewise meta-true.

The map is multilevel, the territory is single-level. This doesn't mean that the higher levels "don't exist", like looking in your garage for a dragon and finding nothing there, or like seeing a mirage in the desert and forming an expectation of drinkable water when there is nothing to drink. The higher levels of your map are not *false,* without referent*;* they have referents *in* the single level of physics. It's not that the wings of an airplane unexist - then the airplane would drop out of the sky. The "wings of an airplane" exist *explicitly* in an engineer's

multilevel model of an airplane, and the wings of an airplane exist *implicitly* in the quantum physics of the real airplane. Implicit existence is not the same as nonexistence. The exact description of this implicitness is not known to us - is not explicitly represented in our map. But this does not prevent our map from working, or even prevent it from being *true.*

Though it is a bit unnerving to contemplate that every single concept and belief in your brain, including these meta-concepts about how your brain works and why you can form accurate beliefs, are perched orders and orders of magnitude above reality. . . *

Zombies! Zombies?

"zombie", in the philosophical usage of the term, is putatively a being that is exactly like you in *every* respect - identical behavior, identical speech, identical brain; every atom and quark in *exactly* the same position, moving according to the same causal laws of motion - *except* that your zombie is not conscious.

It is furthermore claimed that if zombies are "possible" (a term over which battles are still being fought), then, purely from our knowledge of this "possibility", we can deduce a priori that consciousness is extra-physical, in a sense to be described below; the standard term for this position is "epiphenomenalism".

(For those unfamiliar with zombies, I emphasize that *this is not a strawman.* See, for example, the SEP entry on Zombies. The "possibility" of zombies is accepted by a substantial fraction, possibly a majority, of academic philosophers of consciousness.)

I once read somewhere, "You are not the one who speaks your thoughts - you are the one who *hears* your thoughts". In Hebrew, the word for the highest soul, that which God breathed into Adam, is N'Shama - "the hearer".

If you conceive of "consciousness" as a purely passive listening, then the notion of a zombie initially seems easy to imagine. It's someone who lacks the N'Shama, the hearer.

(Warning: Long post ahead. *Very* long 6,600-word post involving David Chalmers ahead. This may be taken as my demonstrative counterexample to Richard Chappell's Arguing with Eliezer Part II, in which Richard accuses me of not engaging with the complex arguments of real philosophers.)

When you open a refrigerator and find that the orange juice is gone, you think "Darn, I'm out of orange juice." The sound of these words is probably represented in your auditory cortex, as though you'd heard someone else say it. (Why do I think this? Because native Chinese speakers can remember longer digit sequences than English-speakers. Chinese digits are all single syllables, and so Chinese speakers can remember around ten digits, versus the famous "seven plus

or minus two" for English speakers. There appears to be a loop of repeating sounds back to yourself, a size limit on working memory in the auditory cortex, which is genuinely phoneme-based.)

Let's suppose the above is correct; as a postulate, it should certainly present no problem for advocates of zombies. Even if humans are not like this, it seems easy enough to imagine an AI constructed this way (and imaginability is what the zombie argument is all about). It's not only conceivable in principle, but quite possible in the next couple of decades, that surgeons will lay a network of neural taps over someone's auditory cortex and read out their internal narrative. (Researchers have already tapped the lateral geniculate nucleus of a cat and reconstructed recognizable visual inputs.)

So your zombie, being physically identical to you down to the last atom, will open the refrigerator and form auditory cortical patterns for the phonemes "Darn, I'm out of orange juice". On this point, epiphenomalists would willingly agree.

But, says the epiphenomenalist, in the zombie there is no one inside to *hear;* the inner listener is missing. The internal narrative is spoken, but unheard. You are not the one who speaks your thoughts, you are the one who hears them.

It seems a lot more straightforward (they would say) to make an AI that prints out some kind of internal narrative, than to show that an inner listener hears it.

The Zombie Argument is that if the Zombie World is*possible* - not necessarily physically possible in our universe, just "possible in theory", or "imaginable", or something along those lines - then consciousness must be extra-physical, something over and above mere atoms. Why? Because even if you somehow knew the positions of all the atoms in the universe, you would still have be told, as a separate and additional fact, that people were conscious - that they had inner listeners - that we were not in the Zombie World, as seems *possible.*

Zombie-ism is not the same as dualism. Descartes thought there was a body-substance and a wholly different kind of mind-substance, but Descartes also thought that the mind-substance was a*causally active* principle, interacting with the body-substance, controlling our speech and behavior. Subtracting out the mind-substance from the human would leave a *traditional* zombie, of the lurching and groaning sort.

And though the Hebrew word for the innermost soul is N'Shama, that-which-hears, I can't recall hearing a rabbi arguing for the possibility of zombies. Most rabbis would probably be aghast at the idea that the divine part which God breathed into Adam *doesn't actually do anything.*

The technical term for the belief that consciousness is there, but has no effect on the physical world, is *epiphenomenalism.*

Though there are other elements to the zombie argument (I'll deal with them below), I think that the intuition of the passive listener is what first seduces

people to zombie-ism. In particular, it's what seduces a lay audience to zombie-ism. The core notion is simple and easy to access: The lights are on but no one's home.

Philosophers are appealing to the intuition of the passive listener when they say "Of course the zombie world is imaginable; you know exactly what it would be like."

One of the great battles in the Zombie Wars is over what, exactly, is meant by saying that zombies are "possible". Early zombie-ist philosophers (the 1970s) just thought it was obvious that zombies were "possible", and didn't bother to define what sort of possibility was meant.

Because of my reading in mathematical logic, what instantly comes into my mind is logical possibility. If you have a collection of statements like (A->B),(B->C),(C->∼A) then the compound belief is *logically possible* if it has a *model* - which, in the simple case above, reduces to finding a value assignment to A, B, C that makes all of the statements (A->B),(B->C), and (C->∼A) true. In this case, A=B=C=0 works, as does A=0, B=C=1 or A=B=0, C=1.

Something will *seem* possible - will seem "conceptually possible" or "imaginable" - if you can consider the collection of statements without *seeing* a contradiction. But it is, in general, a very hard problem to see contradictions *or* to find a full specific model! If you limit yourself to simple Boolean propositions of the form ((A or B or C) and (B or ∼C or D) and (D or ∼A or ∼C) . . .), conjunctions of disjunctions of three variables, then this is a very famous problem called 3-SAT, which is one of the first problems ever to be proven NP-complete.

So just because you don't see a contradiction in the Zombie World at first glance, it doesn't mean that no contradiction is there. It's like not seeing a contradiction in the Riemann Hypothesis at first glance. From conceptual possibility ("I don't see a problem") to *logical possibility* in the full technical sense, is a very great leap. It's easy to make it an NP-complete leap, and with first-order theories you can make it superexponential. And it's *logical* possibility of the Zombie World, not conceptual possibility, that is needed to suppose that a logically omniscient mind could know the positions of all the atoms in the universe, and yet need to be told as an *additional* non-entailed fact that we have inner listeners.

Just because you don't see a contradiction *yet,* is no guarantee that you won't see a contradiction in another 30 seconds. "All odd numbers are prime. Proof: 3 is prime, 5 is prime, 7 is prime. . . "

So let us ponder the Zombie Argument *a little longer*: Can we think of a counterexample to the assertion "Consciousness has no third-party-detectable causal impact on the world"?

If you close your eyes and concentrate on your inward awareness, you will begin to form thoughts, in your internal narrative, that go along the lines of "I am aware" and "My awareness is separate from my thoughts" and "I am not the

one who speaks my thoughts, but the one who hears them" and "My stream of consciousness is not my consciousness" and "It seems like there is a part of me which I can imagine being eliminated without changing my outward behavior."

You can even say these sentences out loud, as you meditate. In principle, someone with a super-fMRI could probably read the phonemes out of your auditory cortex; but saying it out loud removes all doubt about whether you have entered the realms of testability and physical consequences.

This certainly seems like the inner listener is being *caught in the act of listening* by whatever part of you writes the internal narrative and flaps your tongue.

Imagine that a mysterious race of aliens visit you, and leave you a mysterious black box as a gift. You try poking and prodding the black box, but (as far as you can tell) you never succeed in eliciting a reaction. You can't make the black box produce gold coins or answer questions. So you conclude that the black box is causally inactive: "For all X, the black box doesn't do X." The black box is an effect, but not a cause; epiphenomenal; without causal potency. In your mind, you test this general hypothesis to see if it is true in some trial cases, and it seems to be true - "Does the black box turn lead to gold? No. Does the black box boil water? No."

But you can *see* the black box; it absorbs light, and weighs heavy in your hand. This, too, is part of the dance of causality. If the black box were *wholly*outside the causal universe, you couldn't see it; you would have no way to know it existed; you could not say, "Thanks for the black box." You didn't *think* of this counterexample, when you formulated the general rule: "All X: Black box doesn't do X". But it was there all along.

(Actually, the aliens left you *another* black box, this one *purely* epiphenomenal, and you haven't the slightest clue that it's there in your living room. That was their joke.)

If you can close your eyes, and sense yourself sensing - if you can be aware of yourself being aware, and think "I am aware that I am aware" - and say out loud, "I am aware that I am aware" - then your consciousness is not without effect on your internal narrative, or your moving lips. You can see yourself seeing, and your internal narrative reflects this, and so do your lips if you choose to say it out loud.

I have not seen the above argument written out that particular way - "the listener caught in the act of listening" - though it may well have been said before.

But it is a standard point - which zombie-ist philosophers accept! - that the Zombie World's philosophers, being atom-by-atom identical to our own philosophers, write identical papers about the philosophy of consciousness.

At this point, the Zombie World stops being an intuitive consequence of the idea of a passive listener.

Philosophers writing papers about consciousness would *seem* to be at least one effect of consciousness upon the world. You can argue clever reasons why this is not so, but you have to be clever.

You would intuitively suppose that if your inward awareness went away, this would change the world, in that your internal narrative would no longer say things like "There is a mysterious listener within me," because the mysterious listener would be gone. It is usually right *after* you focus your awareness on your awareness, that your internal narrative says "I am aware of my awareness", which suggests that if the first event never happened again, neither would the second. You can argue clever reasons why this is not so, but you have to be clever.

You can form a propositional belief that "Consciousness is without effect", and not *see* any contradiction at first, if you don't realize that talking about consciousness is an effect of being conscious. But once you see the connection from the general rule that consciousness has no effect, to the specific implication that consciousness has no effect on how philosophers write papers about consciousness, zombie-ism stops being intuitive and starts requiring you to postulate strange things.

One strange thing you might postulate is that there's a Zombie Master, a god within the Zombie World who surreptitiously takes control of zombie philosophers and makes them talk and write about consciousness.

A Zombie Master doesn't seem impossible. Human beings often don't sound all that coherent when talking about consciousness. It might not be that hard to fake their discourse, to the standards of, say, a human amateur talking in a bar. Maybe you could take, as a corpus, one thousand human amateurs trying to discuss consciousness; feed them into a non-conscious but sophisticated AI, better than today's models but not self-modifying; and get back discourse about "consciousness" that sounded as sensible as most humans, which is to say, not very.

But this speech about "consciousness" would not be spontaneous. It would not be produced *within*the AI. It would be a recorded imitation of someone else talking. That is just a holodeck, with a central AI writing the speech of the non-player characters. This is *not* what the Zombie World is about.

By supposition, the Zombie World is atom-by-atom identical to our own, except that the inhabitants lack consciousness. Furthermore, the atoms in the Zombie World move under the same laws of physics as in our own world. If there are "bridging laws" that govern *which configurations of atoms evoke consciousness,* those bridging laws are absent. But, by hypothesis, the difference is not experimentally detectable. When it comes to saying whether a quark zigs or zags or exerts a force on nearby quarks - anything experimentally measurable - the same physical laws govern.

The Zombie World has no *room*for a Zombie Master, because a Zombie Master has to control the zombie's lips, and that control is, in principle, experimentally

detectable. The Zombie Master moves lips, therefore it has observable consequences. There would be a point where an electron zags, instead of zigging, because the Zombie Master says so. (Unless the Zombie Master is actually *in* the world, as a pattern of quarks - but then the Zombie World is not atom-by-atom identical to our own, unless you think *this* world also contains a Zombie Master.)

When a philosopher in our world types, "I think the Zombie World is possible", his fingers strike keys in sequence: Z-O-M-B-I-E. There is a chain of causality that can be traced back from these keystrokes: muscles contracting, nerves firing, commands sent down through the spinal cord, from the motor cortex - and then into less understood areas of the brain, where the philosopher's internal narrative first began talking about "consciousness".

And the philosopher's zombie twin strikes the same keys, *for the same reason,* causally speaking. There is no cause within the chain of explanation for why the philosopher writes the way he does, which is not also present in the zombie twin. The zombie twin also has an internal narrative about "consciousness", that a super-fMRI could read out of the auditory cortex. And whatever other thoughts, or other causes of any kind, led to that internal narrative, they are exactly the same in our own universe and in the Zombie World.

So you can't say that the philosopher is writing about consciousness *because of* consciousness, while the zombie twin is writing about consciousness *because of* a Zombie Master or AI chatbot. When you trace back the chain of causality behind the keyboard, to the internal narrative echoed in the auditory cortex, to the cause of the narrative, you must find the *same* physical explanation in our world as in the zombie world.

As the most formidable advocate of zombie-ism, David Chalmers, writes:

> Think of my zombie twin in the universe next door. He talks about conscious experience all the time, in fact, he seems obsessed by it. He spends ridiculous amounts of time hunched over a computer, writing chapter after chapter on the mysteries of consciousness. He often comments on the pleasure he gets from certain sensory qualia, professing a particular love for deep greens and purples. He frequently gets into arguments with zombie materialists, arguing that their position cannot do justice to the realities of conscious experience.
>
> And yet he has no conscious experience at all! In his universe, the materialists are right and he is wrong. Most of his claims about conscious experience are utterly false. But there is certainly a physical or functional explanation of why he makes the claims he makes. After all, his universe is fully law-governed, and no events therein are miraculous, so there must be some explanation of his claims.
>
> . . . Any explanation of my twins behavior will equally count as an explanation of my behavior, as the processes inside his body are precisely mirrored by those inside mine. The explanation of his claims

> obviously does not depend on the existence of consciousness, as there is no consciousness in his world. It follows that the explanation of my claims is also independent of the existence of consciousness.

Chalmers is not arguing *against* zombies; those are his actual beliefs!

> This paradoxical situation is at once delightful and disturbing. It is not obviously fatal to the nonreductive position, but it is at least something that we need to come to grips with...

I would seriously nominate this as the largest bullet ever bitten in the history of time. And that is a backhanded compliment to David Chalmers: A lesser mortal would simply fail to see the implications, or refuse to face them, or rationalize a reason it wasn't so.

Why would anyone bite a bullet that large? Why would anyone postulate unconscious zombies who write papers about consciousness for *exactly the same reason* that our own genuinely conscious philosophers do?

Not because of the first intuition I wrote about, the intuition of the passive listener. That intuition may say that zombies can drive cars or do math or even fall in love, but it doesn't say that zombies write philosophy papers about their passive listeners.

The zombie argument does not rest *solely* on the intuition of the passive listener. If this was all there was to the zombie argument, it would be dead by now, I think. The intuition that the "listener" can be eliminated without effect, would go away as soon as you realized that your internal narrative routinely *seems* to catch the listener in the act of listening.

No, the drive to bite *this* bullet comes from an entirely different intuition - the intuition that no matter how many atoms you add up, no matter how many masses and electrical charges interact with each other, they will never *necessarily* produce a subjective sensation of the mysterious redness of red. It may be a fact about our physical universe (Chalmers says) that putting such-and-such atoms into such-and-such a position, *evokes* a sensation of redness; but if so, it is not a *necessary* fact, it is something to be explained above and beyond the motion of the atoms.

But if you consider the second intuition on its own, without the intuition of the passive listener, it is hard to see why it implies zombie-ism. Maybe there's just a *different kind of stuff,* apart from and additional to atoms, that is *not* causally passive - a soul that actually *does* stuff, a soul that plays a real causal role in why we write about "the mysterious redness of red". Take out the soul, and... well, assuming you just don't fall over in a coma, you certainly won't write any more papers about consciousness!

This is the position taken by Descartes and most other ancient thinkers: The soul is of a different kind, but it *interacts* with the body. Descartes's position is technically known as *substance dualism* - there is a thought-stuff, a mind-stuff, and it is not like atoms; but it is causally potent, interactive, and leaves a visible mark on our universe.

Zombie-ists are *property dualists* - they don't believe in a *separate* soul; they believe that matter in our universe has *additional properties* beyond the physical.

"Beyond the physical"? What does that mean? It means the extra properties are there, but they don't influence the motion of the atoms, like the properties of electrical charge or mass. The extra properties are not experimentally detectable *by third parties; you* know you are conscious, from the *inside*of your extra properties, but no scientist can ever directly detect this from outside.

So the additional properties are there, but not causally active. The extra properties do not move atoms around, which is why they can't be detected by third parties.

And that's why we can (allegedly) imagine a universe just like this one, with all the atoms in the same places, but the extra properties missing, so that everything goes on the same as before, but no one is conscious.

The Zombie World may not be *physically* possible, say the zombie-ists - because it is a fact that all the matter in our universe has the extra properties, or obeys the bridging laws that evoke consciousness - but the Zombie World is *logically* possible: the bridging laws could have been different.

But, once you realize that conceivability is not the same as logical possibility, and that the Zombie World isn't even all that intuitive, why say that the Zombie World is logically possible?

Why, oh why, say that the extra properties are epiphenomenal and indetectable?

We can put this dilemma very sharply: Chalmers believes that there *is* something called consciousness, and this consciousness embodies the true and indescribable substance of the mysterious redness of red. It may be a property beyond mass and charge, but it's *there*, and it *is* consciousness. Now, having said the above, Chalmers furthermore specifies that this true stuff of consciousness is epiphenomenal, without causal potency - but *why say that?*

Why say that you could subtract this true stuff of consciousness, and leave all the atoms in the same place doing the same things? If that's true, we need some *separate* physical explanation for why Chalmers talks about "the mysterious redness of red". That is, there exists both a mysterious redness of red, which is extra-physical, and *an entirely separate*reason,*within*physics, why Chalmers *talks* about the "mysterious redness of red".

Chalmers does confess that these two things seem like they ought to be related, but really, why do you need both? Why not just pick one or the other?

Once you've postulated that there is a mysterious redness of red, why not just say that it interacts with your internal narrative and makes you talk about the "mysterious redness of red"?

Isn't Descartes taking the simpler approach, here? The *strictly* simpler approach?

Why postulate an extramaterial soul, *and then* postulate that the soul has no effect on the physical world, *and then* postulate a mysterious unknown *material* process that causes your internal narrative to talk about conscious experience?

Why not postulate the true stuff of consciousness which no amount of mere mechanical atoms can add up to, *and then,* having gone that far already, let this true stuff of consciousness have causal effects like making philosophers talk about consciousness?

I am not endorsing Descartes's view. But at least I can understand where Descartes is coming from. Consciousness seems mysterious, so you postulate a mysterious stuff of consciousness. Fine.

But now the zombie-ists postulate that this mysterious stuff *doesn't do anything*, so you need a *whole new*explanation for why you *say* you're conscious.

That isn't vitalism. That's something so bizarre that vitalists would spit out their coffee. "When fires burn, they release phlogiston. *But*phlogiston doesn't have any experimentally detectable impact on our universe, so you'll have to go looking for a *separate* explanation of why a fire can melt snow." *What?*

Are property dualists under the impression that if they postulate a new *active*force, something that has a causal impact on observables, they will be sticking their necks out too far?

Me, I'd say that if you postulate a mysterious, separate, additional, inherently mental property of consciousness, above and beyond positions and velocities, then, at that point, you have *already* stuck your neck out as far as it can go. To postulate this stuff of consciousness, and then further postulate that it *doesn't do anything* - for the love of cute kittens, *why?*

There isn't even an obvious career motive. "Hi, I'm a philosopher of consciousness. My subject matter is the most important thing in the universe and I should get lots of funding? Well, it's nice of you to say so, but actually the phenomenon I study doesn't do anything whatsoever." (Argument from career impact is not valid, but I say it to leave a line of retreat.)

Chalmers critiques substance dualism on the grounds that it's hard to see what new theory of physics, what new substance that interacts with matter, could possibly explain consciousness. But property dualism has exactly the same problem. No matter what kind of dual property you talk about, how exactly does it explain consciousness?

When Chalmers postulated an extra property that *is* consciousness, he *took* that

leap across the unexplainable. How does it help his theory to further specify that this extra property *has no effect*? Why not just let it be causal?

If I were going to be unkind, this would be the time to drag in the dragon - to mention Carl Sagan's parable of the dragon in the garage. "I have a dragon in my garage." Great! I want to see it, let's go! "You can't see it - it's an invisible dragon." Oh, I'd like to hear it then. "Sorry, it's an inaudible dragon." I'd like to measure its carbon dioxide output. "It doesn't breathe." I'll toss a bag of flour into the air, to outline its form. "The dragon is permeable to flour."

One motive for trying to make your theory unfalsifiable, is that deep down you fear to put it to the test. Sir Roger Penrose (physicist) and Stuart Hameroff (neurologist) are substance dualists; they think that there is something mysterious going on in quantum, that Everett is wrong and that the "collapse of the wave-function" is physically real, and that this is where consciousness lives and how it exerts causal effect upon your lips when you say aloud "I think therefore I am." Believing this, they predicted that neurons would protect themselves from decoherence long enough to maintain macroscopic quantum states.

This is in the process of being tested, and so far, prospects are not looking good for Penrose -

- but Penrose's basic conduct is scientifically respectable. Not Bayesian, maybe, but still fundamentally healthy. He came up with a wacky hypothesis. He said how to test it. He went out and tried to actually test it.

As I once said to Stuart Hameroff, "I think the hypothesis you're testing is completely hopeless, and your experiments should *definitely* be funded. Even if you don't find exactly what you're looking for, you're looking in a place where no one else is looking, and you might find something interesting."

So a nasty dismissal of epiphenomenalism would be that zombie-ists are afraid to say the consciousness-stuff can have *effects,* because then scientists could go *looking* for the extra properties, and fail to find them.

I don't think this is actually true of Chalmers, though. If Chalmers lacked self-honesty, he could make things a *lot* easier on himself.

(But just in case Chalmers is reading this and does have falsification-fear, I'll point out that if epiphenomenalism is false, then there *is* some other* *explanation for that-which-we-call consciousness, and it will eventually be found, leaving Chalmers's theory in ruins; so if Chalmers cares about his place in history, he has no motive to endorse epiphenomenalism unless he really thinks it's true.)

Chalmers is one of the most frustrating philosophers I know. Sometimes I wonder if he's pulling an "Atheism Conquered". Chalmers does this really *sharp* analysis... and then turns left at the last minute. He lays out everything

that's wrong with the Zombie World scenario, and then, having reduced the whole argument to smithereens, calmly accepts it.

Chalmers does the same thing when he lays out, in calm detail, the problem with saying that our own beliefs in consciousness are justified, when our zombie twins say exactly the same thing for exactly the same reasons and are wrong.

On Chalmers's theory, Chalmers saying that he believes in consciousness cannot be *causally*justified; the belief is not caused by the fact itself. In the absence of consciousness, Chalmers would write the same papers for the same reasons.

On epiphenomenalism, Chalmers saying that he believes in consciousness cannot be justified as the product of a process that systematically outputs true beliefs, because the zombie twin writes the same papers using the same systematic process and is wrong.

Chalmers admits this. Chalmers, in fact, explains the argument in great detail in his book. Okay, so Chalmers has solidly proven that he is not justified in believing in epiphenomenal consciousness, right? No. Chalmers writes:

> Conscious experience lies at the center of our epistemic universe; we have access to it *directly.* This raises the question: what is it that justifies our beliefs about our experiences, if it is not a causal link to those experiences, and if it is not the mechanisms by which the beliefs are formed? I think the answer to this is clear: it is *having* the experiences that justifies the beliefs. For example, the very fact that I have a red experience now provides justification for my belief that I am having a red experience...
>
> Because my zombie twin lacks experiences, he is in a very different epistemic situation from me, and his judgments lack the corresponding justification. It may be tempting to object that if my belief lies in the physical realm, its justification must lie in the physical realm; but this is a *non sequitur.* From the fact that there is no justification in the physical realm, one might conclude that the *physical*portion of me (my brain, say) is not justified in its belief. But the question is whether *I* am justified in the belief, not whether my *brain* is justified in the belief, and if property dualism is correct than there is more to me than my brain.

So - if I've got this thesis right - there's a core you, above and beyond your brain, that believes it is not a zombie, and directly experiences not being a zombie; and so its beliefs are justified.

But Chalmers just *wrote all that stuff down,* in his very physical *book,* and so did the zombie-Chalmers.

The zombie Chalmers can't have written the book *because* of the zombie's core self above the brain; there must be some entirely different reason, within the laws of physics.

It follows that even if there *is* a part of Chalmers hidden away that is conscious and believes in consciousness, directly and without mediation, there is also a *separable subspace* of Chalmers - a causally closed cognitive subsystem that acts entirely *within* physics - and this "outer self" is what speaks Chalmers's internal narrative, and writes papers on consciousness.

I do not see any way to evade the charge that, on Chalmers's own theory, this separable outer Chalmers is deranged. This is the part of Chalmers that is the same in this world, or the Zombie World; and in either world it writes philosophy papers on consciousness *for no valid reason.* Chalmers's philosophy papers are not output by that inner core of awareness and belief-in-awareness, they are output by the mere physics of the internal narrative that makes Chalmers's fingers strike the keys of his computer.

And yet this deranged outer Chalmers is writing philosophy papers that *just happen* to be perfectly right, *by a separate and additional miracle.* Not a logically necessary miracle (then the Zombie World would not be logically possible). A physically contingent miracle, that happens to be true in what we think is our universe, even though science can never distinguish our universe from the Zombie World.

Or at least, that would seem to be the implication of what the self-confessedly deranged outer Chalmers is telling us.

I think I speak for all reductionists when I say Huh?* *

That's not epicycles. That's, "Planetary motions follow these epicycles - but epicycles don't actually *do*anything - there's something else that makes the planets move the same way the epicycles say they should, which I haven't been able to explain - and by the way, I would say this even if there weren't any epicycles."

I have a nonstandard perspective on philosophy because I look at everything with an eye to designing an AI; specifically, a self-improving Artificial General Intelligence with stable motivational structure.

When I think about designing an AI, I ponder principles like probability theory, the Bayesian notion of evidence as differential diagnostic, and above all, reflective coherence. Any self-modifying AI that starts out in a reflectively inconsistent state won't stay that way for long.

If a self-modifying AI looks at a part of itself that concludes "B" on condition A - a part of itself that writes "B" to memory whenever condition A is true - and the AI inspects this part, determines how it (causally) operates in the context of the larger universe, and the AI decides that this part systematically tends to write false data to memory, then the AI has found what appears to be a bug, and the AI will self-modify not to write "B" to the belief pool under condition A.

Any epistemological theory that disregards reflective coherence is not a good theory to use in constructing self-improving AI. This is a knockdown argu-

ment from my perspective, considering what I intend to actually use philosophy *for*. So I have to invent a reflectively coherent theory anyway. And when I do, by golly, reflective coherence turns out to make intuitive sense.

So that's the unusual way in which I tend to think about these things. And now I look back at Chalmers:

The causally closed "outer Chalmers" (that is not influenced in any way by the "inner Chalmers" that has separate additional awareness and beliefs) must be carrying out some systematically unreliable, unwarranted operation which *in some unexplained fashion* causes the internal narrative to produce beliefs about an "inner Chalmers" that are*correct for no logical reason* in what happens to be our universe.

But there's no possible warrant for the outer Chalmers *or any reflectively coherent self-inspecting AI* to believe in this mysterious correctness. A good AI design should, I think, look like a reflectively coherent intelligence embodied in a causal system, with a *testable* theory of how that selfsame causal system produces systematically accurate beliefs on the way to achieving its goals.

So the AI will scan Chalmers and see a closed causal cognitive system producing an internal narrative that is uttering nonsense. Nonsense that seems to have a high impact on what Chalmers thinks *should be considered a morally valuable person.*

This is not a *necessary*problem for Friendly AI theorists. It is *only*a problem if you happen to be an epiphenomenalist. If you believe either the reductionists (consciousness happens *within*the atoms) or the substance dualists (consciousness is *causally potent* immaterial stuff), people talking about consciousness are talking about something real, and a reflectively consistent Bayesian AI can see this by tracing back the chain of causality for what makes people say "consciousness".

According to Chalmers, the causally closed cognitive system of Chalmers's internal narrative is (mysteriously) malfunctioning in a way that, not by necessity, but just in *our* universe, miraculously happens to be correct. Furthermore, the internal narrative asserts "the internal narrative is mysteriously malfunctioning, but miraculously happens to be correctly echoing the justified thoughts of the epiphenomenal inner core", and again, in *our* universe, miraculously happens to be correct.

Oh, come on!

Shouldn't there come a point where you just give up on an idea? Where, on some raw intuitive level, you just go: *What on Earth was I thinking?*

Humanity has accumulated some broad experience with what correct theories of the world look like. *This is not what a correct theory looks like.*

"Argument from incredulity," you say. Fine, you want it spelled out? The said Chalmersian theory postulates multiple unexplained complex miracles. This

drives down its prior probability, by the conjunction rule of probability and Occam's Razor. It is therefore dominated by at least two theories which postulate fewer miracles, namely:

- Substance dualism:
  - There is a stuff of consciousness which is not yet understood, an extraordinary super-physical stuff that *visibly affects* our world; and this stuff is what makes us talk about consciousness.
- Not-quite-faith-based reductionism:
  - That-which-we-name "consciousness" happens *within* physics, in a way not yet understood, just like what happened the last three thousand times humanity ran into something mysterious.
  - Your intuition that no material substance can possibly add up to consciousness is incorrect. If you *actually*knew *exactly* why you talk about consciousness, this would give you new insights, of a form you can't now anticipate; and afterward you would realize that your arguments about normal physics having no room for consciousness were flawed.

Compare to:

- Epiphenomenal property dualism:
  - Matter has additional consciousness-properties which are not yet understood. These properties are epiphenomenal with respect to ordinarily observable physics - they make no difference to the motion of particles.
  - *Separately*, there exists a not-yet-understood reason *within normal physics* why philosophers talk about consciousness and invent theories of dual properties.
  - *Miraculously,* when philosophers talk about consciousness, the bridging laws of *our*world are exactly right to make this talk about consciousness correct, even though it arises from a malfunction (drawing of logically unwarranted conclusions) in the causally closed cognitive system that types philosophy papers.

I know I'm speaking from limited experience, here. But based on my limited experience, the Zombie Argument may be a candidate for *the most deranged idea in all of philosophy.*

There are times when, as a rationalist, you have to believe things that seem weird to you. Relativity seems weird, quantum mechanics seems weird, natural selection seems weird.

But these weirdnesses are pinned down by massive evidence. There's a difference between believing something weird because science has confirmed it overwhelmingly -

- versus believing a proposition that seems downright *deranged,*because of a great big complicated philosophical argument centered around unspecified miracles and giant blank spots not even claimed to be understood -
- in a case where *even if you accept everything that has been told to you so far,* afterward the phenomenon will still seem like a mystery and still have the same quality of wondrous impenetrability that it had at the start.

The correct thing for a rationalist to say at this point, if all of David Chalmers's arguments seem individually plausible - which they don't seem to me - is:

"Okay... I don't know how consciousness works... I admit that... and maybe I'm approaching the whole problem wrong, or asking the wrong questions... but this zombie business *can't possibly be right.* The arguments aren't nailed down enough to make me believe this - especially when accepting it won't make me feel any less confused. On a core gut level, this just *doesn't look* like the way reality could *really really* work."

Mind you, I am not saying this is a substitute for careful analytic refutation of Chalmers's thesis. System 1 is not a substitute for System 2, though it can help point the way. You still have to track down where the problems are *specifically.*

Chalmers wrote a big book, not all of which is available through free Google preview. I haven't duplicated the long chains of argument where Chalmers lays out the arguments against himself in calm detail. I've just tried to tack on a final refutation of Chalmers's last presented defense, which Chalmers has not yet countered to my knowledge. Hit the ball back into his court, as it were.

But, yes, on a core level, the *sane* thing to do when you see the conclusion of the zombie argument, is to say "That can't *possibly*be right" and start looking for a flaw.

Zombie Responses

I'm a bit tired today, having stayed up until 3AM writing yesterday's >6000-word post on zombies, so today I'll just reply to Richard, and tie up a loose end I spotted the next day.

Besides, TypePad's nitwit, un-opt-out-able 50-comment pagination "feature", that doesn't work with the Recent Comments sidebar, means that we might as well jump the discussion here before we go over the 50-comment limit.

(A) Richard Chappell writes:

> A terminological note (to avoid unnecessary confusion): what you call 'conceivable', others of us would merely call "*apparently* conceivable".

The gap between "I don't see a contradiction yet" and "this is logically possible" is so huge (it's NP-complete even in some simple-seeming cases) that you really should have two different words. As the zombie argument is boosted to the extent that this huge gap can be swept under the rug of minor terminological differences, I really think it would be a good idea to say "conceivable" versus "logically possible" or maybe even have a still more visible distinction. I can't choose professional terminology that has already been established, but in a case like this, I might seriously refuse to use it.

Maybe I will say "apparently conceivable" for the kind of information that zombie advocates get by imagining Zombie Worlds, and "logically possible" for the kind of information that is established by exhibiting a complete model or logical proof. Note the size of the gap between the information you can get by closing your eyes and imagining zombies, and the information you need to carry the argument for epiphenomenalism.

> That is, your view would be characterized as a form of Type-A materialism, the view that zombies are not even (genuinely) conceivable, let alone metaphysically possible.

Type-A materialism is a large bundle; you shouldn't attribute the bundle to me until you see me agree with each of the parts. I think that someone who asks "What is consciousness?" is asking a legitimate question, has a legitimate demand for insight; I don't necessarily think that the *answer* takes the form of "Here is this stuff that has all the properties you would attribute to consciousness, for such-and-such reason", but may to some extent consist of insights that cause you to realize you were asking the question the wrong way.

This is not being eliminative about consciousness. It is being realistic about what kind of insights to expect, faced with a problem that (1) seems like it must have *some* solution, (2) seems like it cannot possibly have any solution, and (3) is being *discussed* in a fashion that has a great big dependence on the not-fully-understood ad-hoc architecture of human cognition.

> (1) You haven't, so far as I can tell, identified any *logical contradiction* in the description of the zombie world. You've just pointed out that it's kind of strange. But there are many bizarre possible worlds out there. That's no reason to posit an implicit contradiction. So it's still completely mysterious to me what this alleged contradiction is supposed to be.

Okay, I'll spell it out from a materialist standpoint:

1. The zombie world, by definition, contains all parts of our world that are closed with respect to causality. In particular, it contains the cause of my saying, "I think therefore I am."

2. When I focus my inward awareness on my inward awareness, I shortly thereafter experience my internal narrative saying "I am focusing my inward awareness on my inward awareness", and can, if I choose, say so out loud.

3. Intuitively, it sure seems like my inward awareness is causing my internal narrative to say certain things.

4. The word "consciousness", if it has any meaning at all, refers to that-which-is or that-which-causes or that-which-makes-me-say-I-have inward awareness.

5. From (3) and (4) it would follow that if the zombie world is closed with respect to the causes of my saying "I think therefore I am", the zombie world contains that which we refer to as "consciousness".

6. By definition, the zombie world does not contain consciousness.

7. (3) seems to me to have a rather high probability of being empirically true. Therefore I evaluate a high empirical probability that the zombie world is logically impossible.

You can save the Zombie World by letting the cause of my internal narrative's saying "I think therefore I am" be something entirely other than consciousness. In conjunction with the assumption that consciousness does exist, this is the part that struck me as deranged.

But if the above is *conceivable,* then isn't the Zombie World conceivable?

No, because the two constructions of the Zombie World involve giving the word "consciousness" different empirical referents, like "water" in our world meaning H20 versus "water" in Putnam's Twin Earth meaning XYZ. For the Zombie World to be logically possible, it does not suffice that, for all *you* knew about how the empirical world worked, the word "consciousness" *could* have referred to an epiphenomenon that is entirely different from the consciousness we know. The Zombie World lacks consciousness, not "consciousness" - it is a world without H20, not a world without "water". This is what is required to carry the empirical statement, "You could eliminate the referent of whatever is meant by"consciousness" from our world, while keeping all the atoms in the same place."

Which is to say: I hold that it is an *empirical* fact, given what the word "consciousness" actually refers to, that it is *logically* impossible to eliminate consciousness without moving any atoms. What it would mean to eliminate "consciousness" from a world, rather than consciousness, I will not speculate.

> (2) It's misleading to say it's "miraculous" (on the property dualist view) that our qualia line up so neatly with the physical world. There's a natural law which guarantees this, after all. So it's no more miraculous than any other logically contingent nomic necessity (e.g. the constants in our physical laws).

It is the natural law itself that is "miraculous" - counts as an additional complex-improbable element of the theory to be postulated, without having been itself justified in terms of things already known. One postulates (a) an inner world that is conscious (b) a malfunctioning outer world that talks about consciousness for no reason (c) that the two align perfectly. C does not follow from A and B, and so is a separate postulate.

I agree that this usage of "miraculous" conflicts with the philosophical sense of violating a natural law; I meant it in the sense of improbability appearing from no apparent source, a la perpetual motion belief. Hence the word was ill-chosen in context.

> That is, Zombie (or 'Outer') Chalmers doesn't actually conclude *anything*, because his utterances are meaningless. A fortiori, he doesn't conclude anything unwarrantedly. He's just making noises; these are no more susceptible to epistemic assessment than the chirps of a bird.

Looking at this from an AI-design standpoint, it seems to me like you should be able to build an AI that systematically refines an inner part of itself that correlates (in the sense of mutual information or systematic relations) to the environment, perhaps including floating-point numbers of a sort that I would call "probabilities" because they obey the internal relations mandated by Cox's Theorems when the AI encounters new information - pardon me, new sense inputs.

You will say that, unless the AI is more than mere transistors - unless it has the dual aspect - the AI has no beliefs.

I think my views on this were expressed pretty clearly in "The Simple Truth".

To me, it seems pretty straightforward to construct notions of maps that correlate to territories in systematic ways, without mentioning anything other than things of pure physical causality. The AI outputs a map of Texas. Another AI flies with the map to Texas and checks to see if the highways are in the corresponding places, chirping "True" when it detects a match and "False" when it detects a mismatch. You can refuse to call this "a map of Texas" but the AIs themselves are still chirping "True" or "False", and the said AIs are going to chirp "False" when they look at Chalmers's belief in an epiphenomenal inner core, and I for one would agree with them.

It's clear that the *function of mapping reality* is performed strictly by Outer Chalmers. The whole business of *producing belief representations* is handled by Bayesian structure in causal interactions. There's nothing left for the Inner Chalmers to do, but bless the whole affair with epiphenomenal *meaning.* So when it comes to talking about "accuracy", let alone "systematic accuracy", it seems like we should be able to determine it strictly by looking at the Outer Chalmers.

(**B**) In yesterday's text, I left out an assumption when I wrote:

> If a self-modifying AI looks at a part of itself that concludes "B" on condition A - a part of itself that writes "B" to memory whenever condition A is true - and the AI inspects this part, determines how it (causally) operates in the context of the larger universe, and the AI decides that this part systematically tends to write false data to memory, then the AI has found what appears to be a bug, and the AI will self-modify not to write "B" to the belief pool under condition A...
>
> But there's no possible warrant for the outer Chalmers *or any reflectively coherent self-inspecting AI* to believe in this mysterious correctness. A good AI design should, I think, look like a reflectively coherent intelligence embodied in a causal system, with a *testable* theory of how that selfsame causal system produces systematically accurate beliefs on the way to achieving its goals.

Actually, you need an additional assumption to the above, which is that a "good AI design" (the kind I was thinking of, anyway) judges its own rationality in a modular way; it enforces global rationality by enforcing local rationality. If there is a piece that, relative to its context, is locally systematically unreliable - for some possible beliefs "B_i" and conditions A_i, it adds some "B_i" to the belief pool under local condition A_i, where reflection by the system indicates that B_i is not true (or in the case of probabilistic beliefs, not accurate) when the local condition A_i is true, then this is a bug.

This kind of modularity is *a* way to make the problem tractable, and it's how I currently think about the first-generation AI design, but it may not be the only way to make the problem tractable. Obviously a lot of handwaving here, but you get the general idea.

The notion is that a causally closed cognitive system - such as an AI designed by its programmers to use only causally efficacious parts, or an AI whose theory of its own functioning is entirely testable, or the outer Chalmers that writes philosophy papers - which believes that it has an epiphenomenal inner self, must be doing something systematically unreliable because it would conclude the same thing in a Zombie World. A mind all of whose parts are systematically locally reliable, relative to their contexts, would be systematically globally

reliable. Ergo, a mind which is globally unreliable must contain at least one locally unreliable part. So a causally closed cognitive system inspecting itself for local reliability must discover that at least one step involved in adding the belief of an epiphenomenal inner self, is unreliable.

If there are other ways for minds to be reflectively coherent which avoid this proof of disbelief in zombies, philosophers are welcome to try and specify them.

The reason why I have to specify all this is that otherwise you get a kind of extremely cheap reflective coherence where the AI can never label itself unreliable. E.g. if the AI finds a part of itself that computes $2 + 2 = 5$ (in the surrounding context of counting sheep) the AI will reason: "Well, this part malfunctions and says that $2 + 2 = 5$... but by pure coincidence, $2 + 2$ *is* equal to 5, or so it seems to me... so while the part looks systematically unreliable, I better keep it the way it is, or it will handle this special case wrong." That's why I talk about enforcing global reliability by enforcing local systematic reliability - if you just compare your global beliefs to your global beliefs, you don't go anywhere.

This does have a general lesson: Show your arguments are globally reliable by virtue of each step being locally reliable, don't just compare the arguments' conclusions to your intuitions.

(C) An anonymous poster wrote:

> A sidepoint, this, but I believe your etymology for "n'shama" is wrong. It is related to the word for "breath", not "hear". The root for "hear" contains an ayin, which n'shama does not.

Now that's what I call a miraculously misleading coincidence - although the word N'Shama arose for completely different reasons, it sounded *exactly the right way* to make me think it referred to an inner listener.

Oops.

The Generalized Anti-Zombie Principle

> "Each problem that I solved became a rule which served afterwards to solve other problems."
> — Rene Descartes, *Discours de la Methode*

"Zombies" are putatively beings that are atom-by-atom identical to us, governed by all the same third-party-visible physical laws, except that they are not conscious.

Though the philosophy is complicated, the core argument against zombies is simple: When you focus your inward awareness on your inward awareness,

soon after your internal narrative (the little voice inside your head that speaks your thoughts) says "I am aware of being aware", and then you say it out loud, and then you type it into a computer keyboard, and create a third-party visible blog post.

Consciousness, whatever it may be - a substance, a process, a name for a confusion - is not epiphenomenal; your mind can catch the inner listener in the act of listening, and say so out loud. *The fact that I have typed this paragraph* would at least *seem* to refute the idea that consciousness has no experimentally detectable consequences.

I hate to say "So now let's accept this and move on," over such a philosophically controversial question, but it seems like a considerable majority of Overcoming Bias commenters do accept this. And there are other conclusions you can only get to after you accept that you cannot subtract consciousness and leave the universe looking exactly the same. So now let's accept this and move on.

The form of the Anti-Zombie Argument seems like it should generalize, becoming an Anti-Zombie Principle. But what is the proper generalization?

Let's say, for example, that someone says: "I have a switch in my hand, which does not affect your brain in any way; and iff this switch is flipped, you will cease to be conscious." Does the Anti-Zombie Principle rule this out as well, with the same structure of argument?

It appears to me that in the case above, the answer is yes. In particular, you can say: "Even after your switch is flipped, I will still talk about consciousness *for exactly the same reasons* I did before. If I am conscious right now, I will still be conscious after you flip the switch."

Philosophers may object, "But now you're equating consciousness with talking about consciousness! What about the Zombie Master, the chatbot that regurgitates a remixed corpus of amateur human discourse on consciousness?"

But I did *not* equate "consciousness" with verbal behavior. The core premise is that, *among other things,* the true referent of "consciousness" is *also* the *cause in humans* of talking about inner listeners.

As I argued (at some length) in the sequence on words, what you want in defining a word is not always a perfect Aristotelian necessary-and-sufficient definition; sometimes you just want a treasure map that leads you to the extensional referent. So "that which *does in fact* make me talk about an unspeakable awareness" is not a necessary-and-sufficient definition. But if what does *in fact* cause me to discourse about an unspeakable awareness, is not "consciousness", then...

...then the discourse gets pretty futile. That is not a knockdown argument against zombies - an empirical question can't be settled by mere difficulties of discourse. But if you try to defy the Anti-Zombie Principle, you will have problems with the *meaning* of your discourse, not just its plausibility.

Could we *define* the word "consciousness" to mean "whatever actually makes humans talk about 'consciousness'"? This would have the powerful advantage of guaranteeing that there is at least one real fact named by the word "consciousness". Even if our belief in consciousness is a confusion, "consciousness" would name the cognitive architecture that generated the confusion. But to establish a definition is only to promise to use a word consistently; it doesn't settle any empirical questions, such as whether our inner awareness makes us talk about our inner awareness.

Let's return to the Off-Switch.

If we allow that the Anti-Zombie Argument applies against the Off-Switch, then the Generalized Anti-Zombie Principle does *not* say only, "Any change that is not in-principle experimentally detectable (IPED) cannot remove your consciousness." The switch's flipping is experimentally detectable, but it still seems *highly*unlikely to remove your consciousness.

Perhaps the Anti-Zombie Principle says, "Any change that does not affect you in any IPED way cannot remove your consciousness"?

But is it a reasonable stipulation to say that flipping the switch does not affect you in *any* IPED way? All the particles in the switch are interacting with the particles composing your body and brain. There are gravitational effects - tiny, but real and IPED. The gravitational pull from a one-gram switch ten meters away is around $6 * 10^{-16}$ m/s^2. That's around half a neutron diameter per second per second, far below thermal noise, but way above the Planck level.

We could flip the switch light-years away, in which case the flip would have no immediate causal effect on you (whatever "immediate" means in this case) (if the Standard Model of physics is correct).

But it doesn't seem like we *should* have to alter the thought experiment in this fashion. It seems that, if a disconnected switch is flipped on the other side of a room, you should not expect your inner listener to go out like a light, because the switch "obviously doesn't change" that which is the true cause of your talking about an inner listener. Whatever you really are, you don't expect the switch to mess with it.

This is a *large* step.

If you deny that it is a reasonable step, you had better never go near a switch again. But still, it's a large step.

The key idea of reductionism is that our maps of the universe are multi-level to save on computing power, but physics seems to be strictly single-level. All our discourse about the universe takes place using references far above the level of fundamental particles.

The switch's flip *does* change the fundamental particles of your body and brain. It nudges them by whole neutron diameters away from where they would have otherwise been.

In ordinary life, we gloss a change this small by saying that the switch "doesn't affect you". But it *does* affect you. It changes everything by whole neutron diameters! What could possibly be remaining the same? Only the *description* that you would give of the higher levels of organization - the cells, the proteins, the spikes traveling along a neural axon. As the map is far less detailed than the territory, it must map many different states to the same description.

Any reasonable sort of humanish *description* of the brain that talks about neurons and activity patterns (or even the conformations of individual microtubules making up axons and dendrites) won't change when you flip a switch on the other side of the room. Nuclei are larger than neutrons, atoms are larger than nuclei, and by the time you get up to talking about the *molecular* level, that tiny little gravitational force has vanished from the list of things you bother to *track*.

But if you add up enough tiny little gravitational pulls, they will eventually yank you across the room and tear you apart by tidal forces, so clearly a small effect is *not* "no effect at all".

Maybe the tidal force from that tiny little pull, by an *amazing* coincidence, pulls a single extra calcium ion just a tiny bit closer to an ion channel, causing it to be pulled in just a tiny bit sooner, making a single neuron fire infinitesimally sooner than it would otherwise have done, a difference which amplifies chaotically, finally making a whole neural spike occur that otherwise wouldn't have occurred, sending you off on a different train of thought, that triggers an epileptic fit, that kills you, causing you to cease to be conscious...

If you add up a lot of tiny quantitative effects, you get a big quantitative effect - big enough to mess with anything you care to name. And so claiming that the switch has literally *zero* effect on the things you care about, is taking it too far.

But with just one switch, the force exerted is vastly less than thermal uncertainties, never mind quantum uncertainties. If you don't expect your consciousness to flicker in and out of existence as the result of thermal jiggling, then you certainly shouldn't expect to go out like a light when someone sneezes a kilometer away.

The alert Bayesian will note that I have just made an argument about *expectations,* states of *knowledge,* justified *beliefs* about what can and can't switch off your consciousness.

This doesn't necessarily destroy the Anti-Zombie Argument. Probabilities are not certainties, but the *laws of* probability are theorems; if rationality says you can't believe something on your current information, then that is a law, not a suggestion.

Still, this version of the Anti-Zombie Argument is weaker. It doesn't have the nice, clean, absolutely clear-cut status of, "You can't possibly eliminate consciousness while leaving all the atoms in *exactly* the same place." (Or for

"all the atoms" substitute "all causes with in-principle experimentally detectable effects", and "same wavefunction" for "same place", etc.)

But the new version of the Anti-Zombie Argument still carries. You can say, "I don't know what consciousness really is, and I suspect I may be fundamentally confused about the question. But if the word refers to anything at all, it refers to something that is, among other things, the cause of my talking about consciousness. Now, I don't know why I talk about consciousness. But it happens inside my skull, and I expect it has something to do with neurons firing. Or maybe, if I really understood consciousness, I would have to talk about an even more fundamental level than that, like microtubules, or neurotransmitters diffusing across a synaptic channel. But still, that switch you just flipped has an effect on my neurotransmitters and microtubules that's much, much less than thermal noise at 310 Kelvin. So whatever the true cause of my talking about consciousness may be, I don't expect it to be hugely affected by the gravitational pull from that switch. Maybe it's just a tiny little infinitesimal bit affected? But it's certainly not going to go out like a light. I expect to go on talking about consciousness in *almost exactly* the same way afterward, for *almost exactly* the same reasons."

This application of the Anti-Zombie Principle is weaker. But it's also much more general. And, in terms of sheer common sense, correct.

The reductionist and the substance dualist actually have two different versions of the above statement. The reductionist furthermore says, "Whatever makes me talk about consciousness, it seems likely that the important parts take place on a much higher functional level than atomic nuclei. Someone who understood consciousness could abstract away from individual neurons firing, and talk about high-level cognitive architectures, and still describe how my mind produces thoughts like 'I think therefore I am'. So nudging things around by the diameter of a nucleon, shouldn't affect my consciousness (except maybe with very small probability, or by a very tiny amount, or not until after a significant delay)."

The substance dualist furthermore says, "Whatever makes me talk about consciousness, it's got to be something beyond the computational physics we know, which means that it might very well involve quantum effects. But still, my consciousness doesn't flicker on and off whenever someone sneezes a kilometer away. If it did, I would *notice*. It would be like skipping a few seconds, or coming out of a general anesthetic, or sometimes saying,"I don't think therefore I'm not." So since it's a physical fact that thermal vibrations don't disturb the stuff of my awareness, I don't expect flipping the switch to disturb it either."

Either way, you *shouldn't* expect your sense of awareness to vanish when someone says the word "Abracadabra", even if that does have some infinitesimal physical effect on your brain -

But hold on! If you *hear* someone say the word "Abracadabra", that has a very noticeable effect on your brain - so large, even your brain can notice it. It

may alter your internal narrative; you may think, "Why did that person just say 'Abracadabra'?"

Well, but *still* you expect to go on talking about consciousness in almost exactly the same way afterward, for almost exactly the same reasons.

And again, it's not that "consciousness" is being *equated*to "that which makes you talk about consciousness". It's just that consciousness, *among other things,* makes you talk about consciousness. So anything that makes your consciousness go out like a light, should make you stop talking about consciousness.

If we do something to you, where you don't see how it could *possibly* change your internal narrative - the little voice in your head that sometimes says things like "I think therefore I am", whose words you can choose to say aloud - then it shouldn't make you cease to be conscious.

And this is true even if the internal narrative is just "pretty much the same", and the causes of it are also pretty much the same; among the causes that are pretty much the same, is whatever you mean by "consciousness".

If you're wondering where all this is going, and why it's important to go to such tremendous lengths to ponder such an obvious-seeming Generalized Anti-Zombie Principle, then consider the following debate:

Albert: "Suppose I replaced all the neurons in your head with tiny robotic artificial neurons that had the same connections, the same local input-output behavior, and analogous internal state and learning rules."

Bernice: "That's killing me! There wouldn't be a conscious being there anymore."

Charles: "Well, there'd still be a conscious being there, but it wouldn't be *me*."

Sir Roger Penrose: "The thought experiment you propose is impossible. You *can't* duplicate the behavior of neurons without tapping into quantum gravity. That said, there's not much point in me taking further part in this conversation." *(Wanders away.)*

Albert: "Suppose that the replacement is carried out one neuron at a time, and the swap occurs so fast that it doesn't make any difference to global processing."

Bernice: "How could that possibly be the case?"

Albert: "The little robot swims up to the neuron, surrounds it, scans it, learns to duplicate it, and then suddenly takes over the behavior, between one spike and the next. In fact, the imitation is *so* good, that your outward behavior is just the same as it would be if the brain were left undisturbed. Maybe not *exactly*the same, but the causal impact is much less than thermal noise at 310 Kelvin."

Charles: "So what?"

Albert: "So don't your beliefs violate the Generalized Anti-Zombie Principle? Whatever just happened, it didn't change your internal narrative! You'll go around talking about consciousness for exactly the same reason as before."

Bernice: "Those little robots are a Zombie Master. They'll make me talk about consciousness even though I'm not conscious. The Zombie World is possible if you allow there to be an added, extra, experimentally detectable Zombie Master - which those robots *are*."

Charles: "Oh, that's not right, Bernice. The little robots aren't plotting how to fake consciousness, or processing a corpus of text from human amateurs. They're doing the same thing neurons do, just in silicon instead of carbon."

Albert: "Wait, didn't you just agree with me?"

Charles: "I never said the new person wouldn't be conscious. I said it wouldn't be *me*."

Albert: "Well, obviously the Anti-Zombie Principle generalizes to say that this operation hasn't disturbed the true cause of your talking about this *me* thing."

Charles: "Uh-uh! Your operation certainly did disturb the true cause of my talking about consciousness. It substituted a *different* cause in its place, the robots. Now, just because that new cause *also* happens to be conscious - talks about consciousness for the same *generalized* reason - doesn't mean it's the *same* cause that was originally there."

Albert: "But I wouldn't even have to *tell* you about the robot operation. You wouldn't *notice*. If you think, going on introspective evidence, that you are in an important sense"the same person" that you were five minutes ago, and I do something to you that doesn't change the introspective evidence available to you, then your conclusion that you are the same person that you were five minutes ago should be equally justified. Doesn't the Generalized Anti-Zombie Principle say that if I do something to you that alters your consciousness, let alone makes you a completely different person, then you ought to *notice* somehow?"

Bernice: "Not if you replace me with a Zombie Master. Then there's no one there *to* notice."

Charles: "Introspection isn't perfect. Lots of stuff goes on inside my brain that I don't notice."

Albert: "You're postulating epiphenomenal facts about consciousness and identity!"

Bernice: "No I'm not! I can experimentally detect the difference between neurons and robots."

Charles: "No I'm not! I can experimentally detect the moment when the old me is replaced by a new person."

Albert: "Yeah, and I can detect the switch flipping! You're detecting something that doesn't *make a noticeable difference* to the*true cause* of your talk about consciousness and personal identity. And the proof is, you'll talk just the same way afterward."

Bernice: "That's because of your robotic Zombie Master!"

Charles: "Just because two people talk about 'personal identity' for similar reasons doesn't make them the same person."

I think the Generalized Anti-Zombie Principle supports Albert's position, but the reasons shall have to wait for future posts. I need other prerequisites, and besides, this post is already too long.

But you see the importance of the question, "How far can you generalize the Anti-Zombie Argument and have it still be valid?"

The makeup of future galactic civilizations may be determined by the answer...

Gazp vs Glut

In "The Unimagined Preposterousness of Zombies", Daniel Dennett says:

> To date, several philosophers have told me that they plan to accept my challenge to offer a non-question-begging defense of zombies, but the only one I have seen so far involves postulating a "logically possible" but fantastic being a descendent of Ned Block's Giant Lookup Table fantasy...

A Giant Lookup Table, in programmer's parlance, is when you implement a function as a giant table of inputs and outputs, usually to save on runtime computation. If my program needs to know the multiplicative product of two inputs between 1 and 100, I can write a multiplication algorithm that computes each time the function is called, or I can precompute a Giant Lookup Table with 10,000 entries and two indices. There are times when you *do* want to do this, though not for multiplication - times when you're going to reuse the function a lot and it doesn't have many possible inputs; or when clock cycles are cheap while you're initializing, but very expensive while executing.

Giant Lookup Tables get very large, very fast. A GLUT of all possible twenty-ply conversations with ten words per remark, using only 850-word Basic English, would require $7.6 * 10^{585}$ entries.

Replacing a human brain with a Giant Lookup Table of all possible sense inputs and motor outputs (relative to some fine-grained digitization scheme) would require an *unreasonably large amount* of memory storage. But "in principle", as philosophers are fond of saying, it could be done.

The GLUT is not a zombie in the classic sense, because it is microphysically dissimilar to a human. (In fact, a GLUT can't *really* run on the same physics as a human; it's too large to fit in our universe. For philosophical purposes, we shall ignore this and suppose a supply of unlimited memory storage.)

But is the GLUT a zombie at *all*? That is, does it behave exactly like a human without being conscious?

The GLUT-ed body's tongue talks about consciousness. Its fingers write philosophy papers. In every way, so long as you don't peer inside the skull, the GLUT seems just like a human... which certainly seems like a valid example of a zombie: it behaves just like a human, but there's no one home.

Unless the GLUT is conscious, in which case it wouldn't be a valid example.

I can't recall ever seeing *anyone* claim that a GLUT is conscious. (Admittedly my reading in this area is not up to professional grade; feel free to correct me.) Even people who are accused of being (gasp!) functionalists don't claim that GLUTs can be conscious.

GLUTs are the *reductio ad absurdum* to anyone who suggests that consciousness *is simply* an input-output pattern, thereby disposing of all troublesome worries about what goes on inside.

So what does the Generalized Anti-Zombie Principle (GAZP) say about the Giant Lookup Table (GLUT)?

At first glance, it would seem that a GLUT is the very archetype of a Zombie Master - a distinct, additional, detectable, non-conscious system that animates a zombie and makes it talk about consciousness for *different* reasons.

In the interior of the GLUT, there's merely a very simple computer program that looks up inputs and retrieves outputs. Even talking about a "simple computer program" is overshooting the mark, in a case like this. A GLUT is more like ROM than a CPU. We could equally well talk about a series of switched tracks by which some balls roll out of a previously stored stack and into a trough - *period;* that's *all* the GLUT does.

A spokesperson from People for the Ethical Treatment of Zombies replies: "Oh, that's what all the anti-mechanists say, isn't it? That when you look in the brain, you just find a bunch of neurotransmitters opening ion channels? If ion channels can be conscious, why not levers and balls rolling into bins?"

"The problem isn't the levers," replies the functionalist, "the problem is that a GLUT has the *wrong pattern* of levers. You need levers that implement things like, say, formation of beliefs about beliefs, or self-modeling... Heck, you need the ability to write things to memory just so that time can pass for the computation. Unless you think it's possible to program a conscious being in Haskell."

"I don't know about that," says the PETZ spokesperson, "all I know is that this so-called zombie writes philosophical papers about consciousness. Where

do these philosophy papers come from, if not from consciousness?"

Good question! Let us ponder it deeply.

There's a game in physics called Follow-The-Energy. Richard Feynman's father played it with young Richard:

> It was the kind of thing my father would have talked about: "What makes it go? Everything goes because the sun is shining." And then we would have fun discussing it:
> "No, the toy goes because the spring is wound up," I would say. "How did the spring get wound up?" he would ask.
> "I wound it up."
> "And how did you get moving?"
> "From eating."
> "And food grows only because the sun is shining. So it's because the sun is shining that all these things are moving." That would get the concept across that motion is simply the *transformation* of the sun's power.

When you get a little older, you learn that energy is conserved, never created or destroyed, so the notion of *using up* energy doesn't make much sense. You can never change the total amount of energy, so in what sense are you *using* it?

So when physicists grow up, they learn to play a new game called Follow-The-Negentropy - which is really the same game they were playing all along; only the rules are mathier, the game is more useful, and the principles are harder to wrap your mind around conceptually.

Rationalists learn a game called Follow-The-Improbability, the grownup version of "How Do You Know?" The rule of the rationalist's game is that every improbable-seeming belief needs an equivalent amount of evidence to justify it. (This game has *amazingly similar* rules to Follow-The-Negentropy.)

Whenever someone violates the rules of the rationalist's game, you can find a place in their argument where a quantity of improbability appears from nowhere; and this is as much a sign of a problem as, oh, say, an ingenious design of linked wheels and gears that keeps itself running forever.

The one comes to you and says: "I believe with firm and abiding faith that there's an object in the asteroid belt, one foot across and composed entirely of chocolate cake; you can't prove that this is impossible." But, unless the one had access to some kind of evidence for this belief, it would be highly improbable for a correct belief to form *spontaneously*. So either the one can point to evidence, or the belief won't turn out to be true. "But you can't prove it's *impossible* for my mind to spontaneously generate a belief that happens to be correct!" No, but that kind of spontaneous generation is *highly improbable*, just like, oh, say, an egg unscrambling itself.

In Follow-The-Improbability, it's highly suspicious to even *talk* about a specific hypothesis without having had enough evidence to narrow down the space of possible hypotheses. Why aren't you giving equal air time to a decillion other equally plausible hypotheses? You need sufficient evidence to find the "chocolate cake in the asteroid belt" hypothesis in the hypothesis space - otherwise there's no reason to give it more air time than a trillion other candidates like "There's a wooden dresser in the asteroid belt" or "The Flying Spaghetti Monster threw up on my sneakers."

In Follow-The-Improbability, you are not allowed to pull out big complicated specific hypotheses from thin air without *already* having a corresponding amount of evidence; because it's not realistic to suppose that you could spontaneously start discussing the *true* hypothesis by *pure coincidence.*

A philosopher says, "This zombie's skull contains a Giant Lookup Table of all the inputs and outputs for some human's brain." This is a very *large* improbability. So you ask, "How did this improbable event occur? Where did the GLUT come from?"

Now this is not standard philosophical procedure for thought experiments. In standard philosophical procedure, you are allowed to postulate things like "Suppose you were riding a beam of light..." without worrying about physical possibility, let alone mere improbability. But in this case, the origin of the GLUT matters; and that's why it's important to understand the motivating question, "Where did the improbability come from?"

The obvious answer is that you took a computational specification of a human brain, and used *that* to precompute the Giant Lookup Table. (Thereby creating uncounted googols of human beings, some of them in extreme pain, the supermajority gone quite mad in a universe of chaos where inputs bear no relation to outputs. But damn the ethics, this is for *philosophy.*)

In this case, the GLUT *is* writing papers about consciousness because of a conscious algorithm. The GLUT is no more a zombie, than a cellphone is a zombie because it can talk about consciousness while being just a small consumer electronic device. The cellphone is just transmitting philosophy speeches from whoever happens to be on the other end of the line. A GLUT generated from an originally human brain-specification is doing the same thing.

"All right," says the philosopher, "the GLUT was generated randomly, and *just happens* to have the same input-output relations as some reference human."

How, exactly, did you randomly generate the GLUT?

"We used a true randomness source - a quantum device."

But a quantum device just implements the Branch Both Ways instruction; when you generate a bit from a quantum randomness source, the deterministic result is that one set of universe-branches (locally connected amplitude clouds) see 1, and another set of universes see 0. Do it 4 times, create 16 (sets of) universes.

So, really, this is like saying that you got the GLUT by writing down all possible GLUT-sized sequences of 0s and 1s, in a really damn huge bin of lookup tables; and then reaching into the bin, and *somehow* pulling out a GLUT that happened to correspond to a human brain-specification. Where did the improbability come from?

Because if this *wasn't just a coincidence* - if you had some reach-into-the-bin function that pulled out a human-corresponding GLUT by *design,* not just chance - then that reach-into-the-bin function is probably conscious, and so the GLUT is again a cellphone, not a zombie. It's connected to a human at two removes, instead of one, but it's still a cellphone! Nice try at concealing the source of the improbability there!

Now behold where Follow-The-Improbability has taken us: where is the source of this body's tongue talking about an inner listener? The consciousness isn't in the lookup table. The consciousness isn't in the factory that manufactures lots of possible lookup tables. The consciousness was in whatever *pointed to one particular already-manufactured lookup table,* and said, "Use *that* one!"

You can see why I introduced the game of Follow-The-Improbability. Ordinarily, when we're talking to a person, we tend to think that whatever is inside the skull, must be "where the consciousness is". It's only by playing Follow-The-Improbability that we can realize that the real source of the conversation we're having, is that-which-is-responsible-for the *improbability*of the conversation - however distant in time or space, as the Sun moves a wind-up toy.

"No, no!" says the philosopher. "In the thought experiment, they aren't randomly generating lots of GLUTs, and then using a conscious algorithm to pick out one GLUT that seems humanlike! I am *specifying* that, in this thought experiment,* * they reach into the inconceivably vast GLUT bin, and *by pure chance* pull out a GLUT that is identical to a human brain's inputs and outputs! *There!* I've got you cornered now! You can't play Follow-The-Improbability any further!"

Oh. So your *specification* is the source of the improbability here.

When we play Follow-The-Improbability again, we end up *outside the thought experiment,* looking at the *philosopher.*

That which points to the one GLUT that talks about consciousness, out of all the vast space of possibilities, is now... the conscious person asking us to imagine this whole scenario. And our own brains, which will fill in the blank when we imagine, "What will this GLUT say in response to 'Talk about your inner listener'?"

The moral of this story is that when you follow back discourse about "consciousness", you generally find consciousness. It's not always right in front of you. Sometimes it's very cleverly hidden. But it's there. Hence the Generalized Anti-Zombie Principle.

If there is a Zombie Master in the form of a chatbot that processes and remixes amateur human discourse about "consciousness", the humans who generated the original text corpus are conscious.

If someday you come to understand consciousness, and look back, and see that there's a program you can write which will output confused philosophical discourse that sounds an awful lot like humans without itself being conscious - then when I ask "How did this program come to sound similar to humans?" the answer is that *you* wrote it to sound similar *to conscious humans,* rather than choosing on the criterion of similarity to something else. This doesn't mean your little Zombie Master is conscious - but it does mean I can find consciousness somewhere in the universe by tracing back the chain of causality, which means we're not entirely in the Zombie World.

But suppose someone actually *did* reach into a GLUT-bin and by *genuinely pure chance* pulled out a GLUT that wrote philosophy papers?

Well, then it wouldn't be conscious. IMHO.

I mean, there's got to be more to it than inputs and outputs.

Otherwise even a GLUT would be conscious, right?

Oh, and for those of you wondering how this sort of thing relates to my day job...

In this line of business you meet an awful lot of people who think that an arbitrarily generated powerful AI will be "moral". They can't agree among themselves on why, or what they mean by the word "moral"; but they all agree that doing Friendly AI theory is unnecessary. And when you ask them how an arbitrarily generated AI ends up with moral outputs, they proffer elaborate rationalizations aimed at AIs of that which they deem "moral"; and there are all sorts of problems with this, but the number one problem is, "Are you *sure* the AI would follow the same line of thought you invented to argue human morals, when, unlike you, the AI doesn't start out knowing what *you* want it to rationalize?" You could call the counter-principle Follow-The-Decision-Information, or something along those lines. You can account for an AI that does improbably nice things by telling me how you chose the AI's design from a huge space of possibilities, but otherwise the improbability is being pulled out of nowhere - though more and more heavily disguised, as rationalized premises are rationalized in turn.

So I've already done a whole series of posts which I myself generated using Follow-The-Improbability. But I didn't spell out the rules *explicitly*at that time, because I hadn't done the thermodynamic posts yet...

Just thought I'd mention that. It's amazing how many of my Overcoming Bias posts would coincidentally turn out to include ideas surprisingly relevant to discussion of Friendly AI theory... if you believe in coincidence.

Belief in the Implied Invisible

One generalized lesson *not* to learn from the Anti-Zombie Argument is, "Anything you can't see doesn't exist."

It's tempting to conclude the general rule. It would make the Anti-Zombie Argument much simpler, on future occasions, if we could take this as a premise. But unfortunately that's just not Bayesian.

Suppose I transmit a photon out toward infinity, not aimed at any stars, or any galaxies, pointing it toward one of the great voids between superclusters. Based on standard physics, in other words, I don't expect this photon to intercept anything on its way out. The photon is moving at light speed, so I can't chase after it and capture it again.

If the expansion of the universe is accelerating, as current cosmology holds, there will come a future point where I don't expect to be able to interact with the photon even in principle - a future time beyond which I don't expect the photon's future light cone to intercept my world-line. Even if an alien species captured the photon and rushed back to tell us, they couldn't travel fast enough to make up for the accelerating expansion of the universe.

Should I believe that, in the moment where I can no longer interact with it even in principle, the photon disappears?

No.

It would violate Conservation of Energy. And the second law of thermodynamics. And just about every other law of physics. And probably the Three Laws of Robotics. It would imply the photon knows I care about it and knows exactly when to disappear.

It's a *silly idea*.

But if you can believe in the continued existence of photons that have become experimentally undetectable to you, why doesn't this imply a general license to believe in the invisible?

(If you want to think about this question on your own, do so before the jump. . .)

Though I failed to Google a source, I remember reading that when it was first proposed that the Milky Way was our *galaxy*- that the hazy river of light in the night sky was made up of millions (or even billions) of stars - that Occam's Razor was invoked against the new hypothesis. Because, you see, the hypothesis vastly multiplied the number of "entities" in the believed universe. Or maybe it was the suggestion that "nebulae" - those hazy patches seen through a telescope - might be galaxies full of stars, that got the invocation of Occam's Razor.

Lex parsimoniae: Entia non sunt multiplicanda praeter necessitatem.

That was Occam's original formulation, the law of parsimony: Entities should not be multiplied beyond necessity.

If you postulate billions of stars that no one has ever believed in before, you're multiplying entities, aren't you?

No. There are two Bayesian formalizations of Occam's Razor: Solomonoff Induction, and Minimum Message Length. Neither penalizes galaxies for being big.

Which they had better not do! One of the lessons of history is that what-we-call-reality keeps turning out to be bigger and bigger and huger yet. Remember when the Earth was at the center of the universe? Remember when no one had invented Avogadro's number? If Occam's Razor was weighing against the multiplication of entities every time, we'd have to start doubting Occam's Razor, because it would have consistently turned out to be wrong.

In Solomonoff induction, the complexity of your model is the amount of *code* in the computer program you have to write to simulate your model. The amount of *code,* not the amount of RAM it uses, or the number of cycles it takes to compute. A model of the universe that contains billions of galaxies containing billions of stars, each star made of a billion trillion decillion quarks, will take a lot of RAM to run - but the *code* only has to describe the behavior of the quarks, and the stars and galaxies can be left to run themselves. I am speaking semi-metaphorically here - there are things in the universe besides quarks - but the point is, postulating an extra billion galaxies doesn't count against the size of your code, if you've already described one galaxy. It just takes a bit more RAM, and Occam's Razor doesn't care about RAM.

Why not? The Minimum Message Length formalism, which is nearly equivalent to Solomonoff Induction, may make the principle clearer: If you have to tell someone how your model of the universe works, you don't have to individually specify the location of each quark in each star in each galaxy. You just have to write down some equations. The amount of "stuff" that obeys the equation doesn't affect how long it takes to write the equation down. If you encode the equation into a file, and the file is 100 bits long, then there are 2^{100} other models that would be around the same file size, and you'll need roughly 100 bits of supporting evidence. You've got a limited amount of probability mass; and a priori, you've got to divide that mass up among all the messages you could send; and so postulating a model from within a model space of 2^{100} alternatives, means you've got to accept a 2^{-100} prior probability penalty - but having more galaxies doesn't add to this.

Postulating billions of stars in billions of galaxies doesn't affect the length of your message describing the overall behavior of all those galaxies. So you don't take a probability hit from having the *same*equations describing more things. (So long as your model's predictive successes aren't sensitive to the exact initial conditions. If you've got to specify the exact positions of all the quarks for your model to predict as well as it does, the extra quarks do count as a hit.)

If you suppose that the photon disappears when you are no longer looking at

it, this is an *additional law* in your model of the universe. It's the laws that are "entities", costly under the laws of parsimony. Extra quarks are free.

So does it boil down to, "I believe the photon goes on existing as it wings off to nowhere, because my priors say it's simpler for it to go on existing than to disappear"?

This is what I thought at first, but on reflection, it's not quite right. (And not just because it opens the door to obvious abuses.)

I would boil it down to a distinction between belief in the *implied invisible,* and belief in the *additional invisible.*

When you believe that the photon goes on existing as it wings out to infinity, you're not believing that as an *additional* fact.

What you believe (assign probability to) is a set of simple equations; you believe these equations describe the universe. You believe these equations because they are the simplest equations you could find that describe the evidence. These equations are *highly*experimentally testable; they explain huge mounds of evidence visible in the past, and predict the results of many observations in the future.

You believe these equations, and it is a *logical implication* of these equations that the photon goes on existing as it wings off to nowhere, so you believe that as well.

Your priors, or even your probabilities, don't *directly*talk about the photon. What you assign probability to is not the photon, but the general laws. When you assign probability to the laws of physics as we know them, you *automatically* contribute that same probability to the photon continuing to exist on its way to nowhere - if you believe the logical implications of what you believe.

It's not that you believe in the invisible *as such,* from reasoning about invisible things. Rather the experimental evidence supports certain laws, and belief in those laws logically implies the existence of certain entities that you can't interact with. This is belief in the *implied invisible.*

On the other hand, if you believe that the photon is eaten out of existence by the Flying Spaghetti Monster - maybe on this just one occasion - or even if you believed without reason that the photon hit a dust speck on its way out - then you would be believing in a specific extra invisible event, on its own. If you thought that this sort of thing happened in general, you would believe in a specific extra invisible law. This is belief in the *additional invisible.*

The whole matter would be a lot simpler, admittedly, if we could just rule out the existence of entities we can't interact with, once and for all - have the universe stop existing at the edge of our telescopes. But this requires us to be very silly.

Saying that you shouldn't ever need a separate and additional belief about invisible things - that you only believe invisibles that are *logical implications*

of general laws which are themselves testable, and even then, don't have any further beliefs about them that are not logical implications of visibly testable general rules - actually does seem to rule out all abuses of belief in the invisible, when applied correctly.

Perhaps I should say, "you should assign unaltered prior probability to additional invisibles", rather than saying, "do not believe in them." But if you think of a *belief* as something evidentially additional, something you bother to track, something where you bother to count up support for or against, then it's questionable whether we should ever have additional beliefs about additional invisibles.

There are exotic cases that break this in theory. (E.g: The epiphenomenal demons are watching you, and will torture 3^^^3 victims for a year, somewhere you can't ever verify the event, if you ever say the word "Niblick".) But I can't think of a case where the principle fails in human practice.

Added: To make it clear why you would sometimes want to think about implied invisibles, suppose you're going to launch a spaceship, at nearly the speed of light, toward a faraway supercluster. By the time the spaceship gets there and sets up a colony, the universe's expansion will have accelerated too much for them to ever send a message back. Do you deem it worth the purely altruistic effort to set up this colony, for the sake of all the people who will live there and be happy? Or do you think the spaceship blips out of existence before it gets there? This could be a very real question at some point.

Zombies: The Movie

FADE IN around a serious-looking group of uniformed military officers. At the head of the table, a senior, heavy-set man, GENERAL FRED, speaks.

GENERAL FRED: The reports are confirmed. New York has been overrun... by *zombies*.

COLONEL TODD: Again? But we just had a zombie invasion 28 days ago!

GENERAL FRED: These zombies... are different. They're... *philosophical* zombies.

CAPTAIN MUDD: Are they filled with rage, causing them to bite people?

COLONEL TODD: Do they lose all capacity for reason?

GENERAL FRED: No. They behave... *exactly* like we do... except that they're not conscious.

(*Silence grips the table.*)

COLONEL TODD: Dear God.

GENERAL FRED moves over to a computerized display.

GENERAL FRED: This is New York City, two weeks ago.

The display shows crowds bustling through the streets, people eating in restaurants, a garbage truck hauling away trash.

GENERAL FRED: *This...* is New York City... *now.*

The display changes, showing a crowded subway train, a group of students laughing in a park, and a couple holding hands in the sunlight.

COLONEL TODD: It's worse than I imagined.

CAPTAIN MUDD: How can you tell, exactly?

COLONEL TODD: I've never seen anything so brutally ordinary.

A lab-coated SCIENTIST stands up at the foot of the table.

SCIENTIST: The zombie disease eliminates consciousness without changing the brain in any way. We've been trying to understand how the disease is transmitted. Our conclusion is that, since the disease attacks dual properties of ordinary matter, it must, itself, operate outside our universe. We're dealing with an *epiphenomenal virus.*

GENERAL FRED: Are you sure?

SCIENTIST: As sure as we can be in the total absence of evidence.

GENERAL FRED: All right. Compile a report on every epiphenomenon ever observed. What, where, and who. I want a list of everything that hasn't happened in the last fifty years.

CAPTAIN MUDD: If the virus is epiphenomenal, how do we know it exists?

SCIENTIST: The same way we know *we're* conscious.

CAPTAIN MUDD: Oh, okay.

GENERAL FRED: Have the doctors made any progress on finding an epiphenomenal cure?

SCIENTIST: They've tried every placebo in the book. No dice. Everything they do has an effect.

GENERAL FRED: Have you brought in a homeopath?

SCIENTIST: I tried, sir! I couldn't find any!

GENERAL FRED: Excellent. And the Taoists?

SCIENTIST: They refuse to do anything!

GENERAL FRED: Then we may yet be saved.

COLONEL TODD: What about David Chalmers? Shouldn't he be here?

GENERAL FRED: Chalmers... was one of the first victims.

COLONEL TODD: Oh no.

(*Cut* to the INTERIOR of a cell, completely walled in by reinforced glass, where DAVID CHALMERS paces back and forth.)

DOCTOR: David! David Chalmers! Can you hear me?

CHALMERS: Yes.

NURSE: It's no use, doctor.

CHALMERS: I'm perfectly fine. I've been introspecting on my consciousness, and I can't detect any difference. I *know* I would be expected to say that, but -

The DOCTOR turns away from the glass screen in horror.

DOCTOR: His words, they... they *don't mean anything.*

CHALMERS: This is a grotesque distortion of my philosophical views. This sort of thing can't actually happen!

DOCTOR: Why not?

NURSE: Yes, why not?

CHALMERS: Because -

(*Cut* to two POLICE OFFICERS, guarding a dirt road leading up to the imposing steel gate of a gigantic concrete complex. On their uniforms, a badge reads "BRIDGING LAW ENFORCEMENT AGENCY".)

OFFICER 1: You've got to watch out for those clever bastards. They look like humans. They can talk like humans. They're identical to humans on the atomic level. But they're not human.

OFFICER 2: Scumbags.

The huge noise of a throbbing engine echoes over the hills. Up rides the MAN on a white motorcycle. The MAN is wearing black sunglasses and a black leather business suit with a black leather tie and silver metal boots. His white beard flows in the wind. He pulls to a halt in front of the gate.

The OFFICERS bustle up to the motorcycle.

OFFICER 1: State your business here.

MAN: Is this where you're keeping David Chalmers?

OFFICER 2: What's it to you? You a friend of his?

MAN: Can't say I am. But even zombies have rights.

OFFICER 1: All right, buddy, let's see your qualia.

MAN: I don't have any.

OFFICER 2 suddenly pulls a gun, keeping it trained on the MAN. OFFICER 2: Aha! A zombie!

OFFICER 1: No, zombies claim to have qualia.

OFFICER 2: So he's an ordinary human?

OFFICER 1: No, they also claim to have qualia.

The OFFICERS look at the MAN, who waits calmly.

OFFICER 2: Um...

OFFICER 1: Who *are* you?

MAN: I'm Daniel Dennett, bitches.

Seemingly from nowhere, DENNETT pulls a sword and slices OFFICER 2's gun in half with a steely noise. OFFICER 1 begins to reach for his own gun, but DENNETT is suddenly standing behind OFFICER 1 and chops with a fist, striking the junction of OFFICER 1's shoulder and neck. OFFICER 1 drops to the ground.

OFFICER 2 steps back, horrified.

OFFICER 2: That's not possible! How'd you do that?

DENNETT: I am one with my body.

DENNETT drops OFFICER 2 with another blow, and strides toward the gate. He looks up at the imposing concrete complex, and grips his sword tighter.

DENNETT *(quietly to himself)*: There is a spoon.

(*Cut* back to GENERAL FRED and the other military officials.)

GENERAL FRED: I've just received the reports. We've lost Detroit.

CAPTAIN MUDD: I don't want to be the one to say "Good riddance", but -

GENERAL FRED: Australia has been... *reduced to atoms.*

COLONEL TODD: The epiphenomenal virus is spreading faster. Civilization itself threatens to dissolve into total normality. We could be looking at the middle of humanity.

CAPTAIN MUDD: Can we negotiate with the zombies?

GENERAL FRED: We've sent them messages. They sent only a single reply.

CAPTAIN MUDD: Which was...?

GENERAL FRED: It's on its way now.

An orderly brings in an envelope, and hands it to GENERAL FRED.

GENERAL FRED opens the envelope, takes out a single sheet of paper, and reads it.

Silence envelops the room.

CAPTAIN MUDD: What's it say?

GENERAL FRED: It says... that *we're* the ones with the virus.

(A silence falls.)

COLONEL TODD raises his hands and stares at them.

COLONEL TODD: My God, it's true. It's true. I...

(A tear rolls down COLONEL TODD's cheek.)

COLONEL TODD: I don't feel anything.

The screen goes black.

The sound goes silent.

The movie continues exactly as before.

PS: This is me being attacked by zombie nurses at Penguicon.

Only at a *combination*science fiction and open-source convention would it be possible to attend a session on knife-throwing, cry "In the name of Bayes, die!", throw the knife, and then have a fellow holding a wooden shield say, "Yes, but how do you determine the prior for where the knife hits?"

Excluding the Supernatural

Occasionally, you hear someone claiming that creationism should not be taught in schools, especially not as a competing hypothesis to evolution, because creationism is *a priori and automatically* excluded from scientific consideration, in that it invokes the "supernatural".

So... is the idea here, that creationism *could* be true, but *even if it were true*, you wouldn't be *allowed* to teach it in science class, because science is only about "natural" things?

It seems clear enough that this notion stems from the desire to avoid a confrontation between science and religion. You don't want to come right out and say that science doesn't teach Religious Claim X because X has been tested by the scientific method and found false. So instead, you can... um... claim that science is excluding hypothesis X *a priori*. That way you don't have to discuss how experiment has falsified X *a posteriori*.

Of course this plays right into the creationist claim that Intelligent Design isn't getting a fair shake from science - that science has *prejudged* the issue in favor

of atheism, regardless of the evidence. If science excluded Intelligent Design *a priori,* this would be a justified complaint!

But let's back up a moment. The one comes to you and says: "Intelligent Design is excluded from being science *a priori,* because it is 'supernatural', and science only deals in 'natural' explanations."

What exactly do they mean, "supernatural"? Is any explanation invented by someone with the last name "Cohen" a supernatural one? If we're going to summarily kick a set of hypotheses out of science, what is it that we're supposed to exclude?

By *far* the best definition I've ever heard of the supernatural is Richard Carrier's: A "supernatural" explanation appeals to *ontologically basic mental things,* mental entities that cannot be reduced to nonmental entities.

This is the difference, for example, between saying that water rolls downhill because it *wants* to be lower, and setting forth differential equations that claim to describe only motions, not desires. It's the difference between saying that a tree puts forth leaves because of a tree spirit, versus examining plant biochemistry. Cognitive science takes the fight against supernaturalism into the realm of the mind.

Why is this an excellent definition of the supernatural? I refer you to Richard Carrier for the full argument. But consider: Suppose that you discover what seems to be a *spirit,* inhabiting a tree: a dryad who can materialize outside or inside the tree, who speaks in English about the need to protect her tree, et cetera. And then suppose that we turn a microscope on this tree spirit, and she turns out to be made of parts - not inherently spiritual and ineffable parts, like fabric of desireness and cloth of belief; but rather the same sort of parts as quarks and electrons, parts whose behavior is defined in motions rather than minds. Wouldn't the dryad immediately be demoted to the dull catalogue of common things?

But if we accept Richard Carrier's definition of the supernatural, then a dilemma arises: we *want* to give religious claims a fair shake, but it seems that we have *very good* grounds for excluding supernatural explanations *a priori.*

I mean, what *would* the universe look like if reductionism were false?

I previously defined the reductionist thesis as follows: human minds create multi-level *models* of reality in which high-level patterns and low-level patterns are separately and explicitly *represented.* A physicist knows Newton's equation for gravity, Einstein's equation for gravity, and the derivation of the former as a low-speed approximation of the latter. But these three separate mental representations, are only a convenience of human cognition. It is not that *reality itself* has an Einstein equation that governs at high speeds, a Newton equation that governs at low speeds, and a "bridging law" that smooths the interface. Reality itself has only a single level, Einsteinian gravity. It is only the Mind Projection Fallacy that makes some people talk as if the higher levels could

have a separate existence - different levels of organization can have separate representations in human maps, but the territory itself is a single unified low-level mathematical object.

Suppose this were wrong.

Suppose that the Mind Projection Fallacy was not a fallacy, but simply true.

Suppose that a 747 had a fundamental physical existence apart from the quarks making up the 747.

What experimental observations would you expect to make, if you found yourself in such a universe?

If you can't come up with a good answer to that, it's not *observation* that's ruling out "non-reductionist" beliefs, but *a priori* logical incoherence. If you can't say what predictions the "non-reductionist" model makes, how can you say that experimental evidence rules it out?

My thesis is that non-reductionism is a *confusion;* and once you realize that an idea is a confusion, it becomes a tad difficult to envision what the universe would look like if the confusion were *true.* Maybe I've got some multi-level model of the world, and the multi-level model has a one-to-one direct correspondence with the causal elements of the physics? But once all the rules are specified, why wouldn't the model just flatten out into yet another list of fundamental things and their interactions? Does everything I can *see in* the model, like a 747 or a human mind, have to become a separate real thing? But what if I see a pattern in that new supersystem?

Supernaturalism is a special case of non-reductionism, where it is not 747s that are irreducible, but just (some) mental things. Religion is a special case of supernaturalism, where the irreducible mental things are God(s) and souls; and perhaps also sins, angels, karma, etc.

If I propose the existence of a powerful entity with the ability to survey and alter each element of our observed universe, but with the entity reducible to nonmental parts that interact with the elements of our universe in a lawful way; if I propose that this entity wants certain particular things, but "wants" using a brain composed of particles and fields; then this is not yet a religion, just a naturalistic hypothesis about a naturalistic Matrix. If tomorrow the clouds parted and a vast glowing amorphous figure thundered forth the above description of reality, then this would not imply that the figure was necessarily honest; but I would show the movies in a science class, and I would try to derive testable predictions from the theory.

Conversely, religions have ignored the discovery of that ancient bodiless thing: omnipresent in the working of Nature and immanent in every falling leaf: vast as a planet's surface and billions of years old: itself unmade and arising from the structure of physics: designing without brain to shape all life on Earth and the minds of humanity. Natural selection, when Darwin proposed it, was not hailed as the long-awaited Creator: It wasn't *fundamentally* mental.

But now we get to the dilemma: if the staid conventional normal boring understanding of physics and the brain *is* correct, there's no way *in principle* that a human being can concretely envision, and derive testable experimental predictions about, an alternate universe in which things *are* irreducibly mental. Because, if the boring old normal model is correct, your brain is made of quarks, and so your brain will only be able to envision and concretely predict things that can predicted by quarks. You will only ever be able to construct models made of interacting simple things.

People who live in reductionist universes cannot concretely envision non-reductionist universes. They can pronounce the syllables "non-reductionist" but they can't *imagine* it.

The basic error of anthropomorphism, and the reason why supernatural explanations sound much simpler than they really are, is your brain using itself as an opaque black box to predict other things labeled "mindful". Because you already have big, complicated webs of neural circuitry that implement your "wanting" things, it seems like you can easily describe water that "wants" to flow downhill - the one word "want" acts as a lever to set your *own* complicated wanting-machinery in motion.

Or you imagine that God likes beautiful things, and therefore made the flowers. Your own "beauty" circuitry determines what is "beautiful" and "not beautiful". But you don't know the diagram of your own synapses. You can't describe a *nonmental* system that computes the same label for what is "beautiful" or "not beautiful" - can't write a computer program that predicts your own labelings. But this is just a defect of knowledge on your part; it doesn't mean that the brain has no explanation.

If the "boring view" of reality is correct, then you can *never* predict anything irreducible because *you* are reducible. You can never get Bayesian confirmation for a hypothesis of irreducibility, because any *prediction you can make* is, therefore, something that could also be predicted by a reducible thing, namely your brain.

Some boxes you really *can't* think outside. If our universe *really is* Turing computable, we will never be able to *concretely* envision anything that isn't Turing-computable - no matter how many levels of halting oracle hierarchy our mathematicians can talk *about,* we won't be able to predict what a halting oracle would actually *say,* in such fashion as to experimentally discriminate it from merely computable reasoning.

Of course, that's all assuming the "boring view" is correct. *To the extent* that you believe evolution is true, you should not expect to encounter strong evidence against evolution. To the extent you believe reductionism is true, you should expect non-reductionist hypotheses to be *incoherent* as well as wrong. To the extent you believe supernaturalism is false, you should expect it to be *inconceivable* as well.

If, on the other hand, a supernatural hypothesis turns out to be true, then presumably you will also discover that it is not inconceivable.

So let us bring this back full circle to the matter of Intelligent Design:

Should ID be excluded *a priori* from experimental falsification and science classrooms, because, by invoking the supernatural, it has placed itself outside of natural philosophy?

I answer: "Of course not." The *irreducibility* of the intelligent designer is not an indispensable part of the ID hypothesis. For every irreducible God that can be proposed by the IDers, there exists a corresponding reducible alien that behaves in accordance with the same predictions - since the IDers themselves are reducible; to the extent I believe reductionism is in fact correct, which is a rather strong extent, I must expect to discover reducible formulations of all supposedly supernatural predictive models.

If we're going over the archeological records to test the assertion that Jehovah parted the Red Sea out of an explicit desire to display its superhuman power, then it makes little difference whether Jehovah is ontologically basic, or an alien with nanotech, or a Dark Lord of the Matrix. You do some archeology, find no skeletal remnants or armor at the Red Sea site, and indeed find records that Egypt ruled much of Canaan at the time. So you stamp the historical record in the Bible "disproven" and carry on. The hypothesis is coherent, falsifiable and wrong.

Likewise with the evidence from biology that foxes are designed to chase rabbits, rabbits are designed to evade foxes, and neither is designed "to carry on their species" or "protect the harmony of Nature"; likewise with the retina being designed backwards with the light-sensitive parts at the bottom; and so on through a thousand other items of evidence for splintered, immoral, incompetent design. The Jehovah model of our alien god is coherent, falsifiable, and wrong - coherent, that is, so long as you don't care whether Jehovah is ontologically basic or just an alien.

Just convert the supernatural hypothesis into the corresponding natural hypothesis. Just make the same predictions the same way, without asserting any mental things to be ontologically basic. Consult your brain's black box if necessary to make predictions - say, if you want to talk about an "angry god" without building a full-fledged angry AI to label behaviors as angry or not angry. So you derive the predictions, or look up the predictions made by ancient theologians without advance knowledge of our experimental results. If experiment conflicts with those predictions, then it is fair to speak of the religious claim having been scientifically refuted. It was given its just chance at confirmation; it is being excluded *a posteriori,* not *a priori.*

Ultimately, reductionism is just disbelief in *fundamentally complicated* things. If "fundamentally complicated" sounds like an oxymoron... well, that's why I think that the doctrine of non-reductionism is a *confusion,* rather than a way

that things could be, but aren't. You would be wise to be wary, if you find yourself supposing such things.

But the ultimate rule of science is to look and see. If ever a God appeared to thunder upon the mountains, it would be something that people looked at and saw.

Corollary: Any supposed designer of Artificial General Intelligence who talks about religious beliefs in respectful tones, is clearly not an expert on reducing mental things to nonmental things; and indeed knows so very little of the uttermost basics, as for it to be scarcely plausible that they could be expert at the art; unless their *idiot savancy* is complete. Or, of course, if they're outright lying. We're not talking about a subtle mistake.

Psychic Powers

Yesterday, I wrote:

> If the "boring view" of reality is correct, then you can *never* predict anything irreducible because *you* are reducible. You can never get Bayesian confirmation for a hypothesis of irreducibility, because any *prediction you can make* is, therefore, something that could also be predicted by a reducible thing, namely your brain.

Benja Fallenstein commented:

> I think that while you can in this case never devise an empirical test whose outcome could *logically prove* irreducibility, there is no clear reason to believe that you cannot devise a test whose counterfactual outcome in an irreducible world would make irreducibility subjectively *much more probable* (given an Occamian prior).
>
> Without getting into reducibility/irreducibility, consider the scenario that the physical universe makes it possible to build a hypercomputer — that performs operations on arbitrary real numbers, for example — but that our brains do not actually make use of this: they can be simulated perfectly well by an ordinary Turing machine, thank you very much...

Well, that's a very intelligent argument, Benja Fallenstein. But I have a crushing reply to your argument, such that, once I deliver it, you will at once give up further debate with me on this particular point:

You're right.

Alas, I don't get modesty credit on this one, because after publishing yesterday's post I realized a similar flaw on my own - this one concerning Occam's Razor and psychic powers:

If beliefs and desires are irreducible and ontologically basic entities, or have an ontologically basic *component* not covered by existing science, that would make it far more likely that there was an ontological rule governing the interaction of different minds - an interaction which bypassed ordinary "material" means of communication like sound waves, known to existing science.

If naturalism is correct, then there exists a conjugate reductionist model that makes the *same predictions* as any concrete prediction that any parapsychologist can make about telepathy.

Indeed, if naturalism is correct, the only reason we can *conceive* of beliefs as "fundamental" is due to lack of self-knowledge of our own neurons - that the peculiar reflective architecture of our own minds exposes the "belief" classbut hides the machinery behind it.

Nonetheless, the discovery of information transfer between brains, in the absence of any known material connection between them, is *probabilistically* a privileged prediction of supernatural models (those that contain ontologically basic mental entities). Just because it is so much *simpler* in that case to have a new law relating beliefs between different minds, compared to the "boring" model where beliefs are complex constructs of neurons.

The hope of psychic powers arises from treating beliefs and desires as sufficiently fundamental objects that they can have *unmediated* connections to reality. If beliefs are patterns of neurons made of known material, with inputs given by organs like eyes constructed of known material, and with outputs through muscles constructed of known material, and this seems sufficient to account for all known mental powers of humans, then there's no reason to expect anything more - no reason to postulate additional connections. This is why reductionists don't expect psychic powers. Thus, observing psychic powers would be strong evidence for the supernatural in Richard Carrier's sense.

We have an Occam rule that counts the number of ontologically basic classes and ontologically basic laws in the model, and penalizes the count of entities. If naturalism is correct, then the attempt to count "belief" or the "relation between belief and reality" as a single basic entity, is simply misguided anthropomorphism; we are only tempted to it by a quirk of our brain's internal architecture. But if you *just go with* that misguided view, then it assigns a much higher probability to psychic powers than does naturalism, because you can implement psychic powers using apparently simpler laws.

Hence the actual discovery of psychic powers would imply that the human-naive Occam rule was in fact better-calibrated than the sophisticated naturalistic Occam rule. It would argue that reductionists had been wrong all along in trying to take apart the brain; that what our minds exposed as a seemingly

simple lever, was in fact a simple lever. The naive dualists would have been right from the beginning, which is why their ancient wish would have been enabled to come true.

So telepathy, and the ability to influence events just by wishing at them, and precognition, would all, if discovered, be strong Bayesian evidence in favor of the hypothesis that beliefs are ontologically fundamental. Not logical proof, but strong Bayesian evidence.

If reductionism is correct, then any science-fiction story containing psychic powers, can be output by a system of simple elements (i.e., the story's author's brain); but if we *in fact* discover psychic powers, that would make it much more probable that events were occurring which could not *in fact* be described by reductionist models.

Which just goes to say: The existence of psychic powers is a privileged probabilistic assertion of non-reductionist worldviews - *they own* that advance prediction; they devised it and put it forth, in defiance of reductionist expectations. So by the laws of science, if psychic powers are discovered, non-reductionism wins.

I am therefore confident in dismissing psychic powers as *a priori* implausible, despite all the claimed experimental evidence in favor of them.