MIRI
MACHINE INTELLIGENCE
RESEARCH INSTITUTE

A Technical Explanation of Technical Explanation

Eliezer Yudkowsky
Machine Intelligence Research Institute

Yudkowsky, Eliezer. 2005. "A Technical Explanation of Technical Explanation."
Unpublished manuscript. `http://yudkowsky.net/rational/technical`

This version contains minor changes.

This essay is meant for a reader who has attained a firm grasp of Bayes' Theorem. An introduction to Bayes' Theorem may be found in *An Intuitive Explanation of Bayesian Reasoning* (Yudkowsky 2003). You should easily recognize, and intuitively understand, the concepts *prior probability*, *posterior probability*, *likelihood ratio*, and *odds ratio*. This essay is intended as a sequel to the *Intuitive Explanation*, but you might skip that introduction if you are already thoroughly Bayesian. Where the *Intuitive Explanation* focused on providing a firm grasp of Bayesian basics, *A Technical Explanation of Technical Explanation* builds, on a Bayesian foundation, theses about human rationality and philosophy of science.

The *Intuitive Explanation of Bayesian Reasoning* promised that mastery of addition, multiplication, and division would be sufficient background, with no subtraction required. To this the *Technical Explanation of Technical Explanation* adds logarithms. The math is simple, but necessary, and it appears first in the order of exposition. Some pictures may not be drawn with words alone.

As Jaynes (2003) emphasizes, the theorems of Bayesian probability theory are just that, *mathematical theorems* which follow inevitably from Bayesian axioms. One might naively think that there would be no controversy about mathematical theorems. But when do the theorems apply? How do we use the theorems in real-world problems? The *Intuitive Explanation* tries to avoid controversy, but the *Technical Explanation* willfully walks into the whirling helicopter blades. Bluntly, the reasoning in the *Technical Explanation* does not represent the unanimous consensus of Earth's entire planetary community of Bayesian researchers. At least, not yet.

The *Technical Explanation of Technical Explanation* is so named because it begins with this question:

> "What is the difference between a technical understanding and a verbal understanding?"

* * *

A fable:

> Once upon a time, there was a teacher who cared for a group of physics students. One day she called them into her class, and showed them a wide, square plate of metal, next to a hot radiator. The students each put their hand on the plate, and found the side next to the radiator cool, and the distant side warm. And the teacher said, write down your guess why this happens. Some students guessed convection of air currents, and others guessed strange patterns of metals in the plate, and not one put down "This seems to me impossible," and the answer was that before the students entered the room, the teacher turned the plate around. (Taken from Verhagen 2001)

There are many morals to this fable, and I have told it with different morals in different contexts. I usually take the moral that your strength as a rationalist is measured by your ability to be more confused by fiction than by reality. If you are equally good at explaining any story, you have zero knowledge. Occasionally I have heard a story that sounds confusing, and reflexively suppressed my feeling of confusion and accepted the story, and then later learned that the original story was untrue. Each time this happens to me, I vow anew to focus consciously on my fleeting feelings of bewilderment.

But in this case, the moral is that the apocryphal students failed to understand what constituted a scientific explanation. If the students measured the heat of the plate at different points and different times, they would soon see a pattern in the numbers. If the students knew the diffusion equation for heat, they might calculate that the plate equilibrated with the radiator and environment two minutes and fifteen seconds ago, turned around, and now approaches equilibrium again. Instead the students wrote down words on paper, and thought they were doing physics. I should rather compare it to the random guessing of Greek philosophers, such as Heraclitus who said "All is Fire," and fancied it his theory of everything.

As a child I read books of popular physics, and fancied myself knowledgeable; I knew sound was waves of air, light was waves of electromagnetism, matter was waves of complex probability amplitudes. When I grew up I read the *Feynman Lectures on Physics* (Feynman, Leighton, and Sands 1963), and discovered a gem called "the wave equation." I thought about that equation, on and off for three days, until I saw to my satisfaction it was dumbfoundingly simple. And when I understood, I realized that during all the time I had believed the honest assurance of physicists that sound and light and matter were waves, I had not the vaguest idea what "wave" meant to a physicist.

* * *

So that is the difference between a technical understanding and a verbal understanding.

Do you believe that? If so, you should have applied the knowledge, and said: "But why didn't you give a technical explanation instead of a verbal explanation?"

* * *

In *An Intuitive Explanation of Bayesian Reasoning* I tried to provide visual and physical metaphors for Bayesian probability; for example, evidence is a *weight*, a *pressure* upon belief, that *slides* prior probabilities to posterior probabilities.

Now we add a new metaphor, which is also the mathematical terminology: Visualize *probability density* or *probability mass*—probability as a lump of clay that you must distribute over possible outcomes.

Let's say there's a little light that can flash *red*, *blue*, or *green* each time you press a button. The light flashes one and only one color on each press of the button; the

possibilities are mutually exclusive. You're trying to predict the color of the next flash. On each try, you have a weight of clay, the probability mass, that you have to distribute over the possibilities red, green, and blue. You might put a fourth of your clay on the green possibility, a fourth of your clay on the blue possibility, and half your clay on the red possibility—like assigning probabilities of 25% to green, 25% to blue, and 50% to red. The metaphor is that *probability is a conserved resource*, to dole out sparingly. If you think that blue is more likely to flash on the next experiment, you can assign a higher probability to blue, but you have to take the probability mass from the other hypotheses—maybe steal some clay from red and add it to blue. You can never get any more clay. Your probabilities can't sum to more than 1.0 (100%). You can't predict a 75% chance of seeing red and an 80% chance of seeing blue.

Why would you want to be careful with your probability mass, or dole it out sparingly? Why not slop probability all over the place? Let's shift the metaphor from clay to money. You can bet up to a dollar of play money on each press of the button. An experimenter stands nearby, and pays you an amount of real money that depends on how much play money you bet on the *winning* light. We don't care how you distributed your remaining play money over the losing lights. The only thing that matters is how much you bet on the light that actually won.

But we must carefully construct the scoring rule used to pay off the winners, if we want the players to be careful with their bets. Suppose the experimenter pays each player real money equal to the play money bet on the winning color. Under this scoring rule, if you observe that red comes up six times out of ten, your best strategy is to bet, not 60 cents on red, but the entire dollar on red, and you don't care about the frequencies of blue and green. Why? Let's say that blue and green each come up around two times out of ten. And suppose you bet 60 cents on red, 20 cents on blue, and 20 cents on green. In this case, six times out of ten you would win 60 cents, and four times out of ten you would win 20 cents, for an average payoff of 44 cents. Under that scoring rule, it makes more sense to allocate the entire dollar to red, and win an entire dollar six times out of ten. Four times out of ten you would win nothing. Your average payoff would be 60 cents.

If we wrote down the function for the payoff, it would be $\text{Payoff} = P(\text{winner})$, where $P(\text{winner})$ is the amount of play money you bet on the winning color on that round. If we wrote down the function for the expected payoff given that Payoff rule, it would be:

$$\text{Expectation}(\text{Payoff}) = \text{Sum}[P(\text{color}) \times F(\text{color})] \text{ for each color.}$$

$P(\text{color})$ is the amount of play money you bet on a color, and $F(\text{color})$ is the frequency with which that color wins.

Suppose that the actual frequencies of the lights are 30% blue, 20% green, and 50% red. And suppose that on each round I bet 40% on blue, 50% on green, and 10% on red.

I would get 40 cents 30% of the time, 50 cents 20% of the time, and 10 cents 50% of the time, for an average payoff of $0.12 + $0.10 + $0.05 or $0.27. That is:

$$P(\text{color}) = \text{play money assigned to that color}$$
$$F(\text{color}) = \text{frequency with which that color wins}$$
$$\text{Payoff} = P(\text{winner}) = \text{amount of play money allocated to winning color}$$

Actual frequencies of winning:

$$F(\text{blue}) = 30\%$$
$$F(\text{green}) = 20\%$$
$$F(\text{red}) = 50\%$$

In the long run, red wins 50% of the time, green wins 20% of the time, and blue wins 30% of the time. So our *average* payoff on each round is 50% of the payoff if red wins, plus 20% of the payoff if green wins, plus 30% of the payoff if blue wins.

The payoff is a function of the winning color and the betting scheme. We want to compute the *average* payoff, given a betting scheme and the *frequencies* at which each color wins. The mathematical term for this kind of computation, taking a function of each case and weighting it by the frequency of that case, is an *expectation*. Thus, to compute our *expected payoff* we would calculate:

$$\begin{aligned}\text{Expectation(Payoff)} &= \text{Sum}[P(\text{color}) \times F(\text{color})] \text{ for each color} \\ &= P(\text{blue}) \times F(\text{blue}) + P(\text{green}) \times F(\text{green}) + P(\text{red}) \times F(\text{red}) \\ &= \$0.40 \times 30\% + \$0.50 \times 20\% + \$0.10 \times 50\% \\ &= \$0.12 + \$0.10 + \$0.05 \\ &= \$0.27\end{aligned}$$

With this betting scheme I'll win, on average, around 27 cents per round.

I allocated my play money in a grossly arbitrary way, and the question arises: Can I increase my expected payoff by allocating my play money more wisely? *Given the scoring rule provided*, I maximize my expected payoff by allocating my *entire* dollar to red. Despite my *expected* payoff of 50 cents per round, the light might *actually* flash green, blue, blue, green, green and I would receive an *actual* payoff of zero. However, the chance of the light coming up non-red on five successive rounds is approximately 3%.

Tversky and Edwards (1966) conducted an experiment.[1] Subjects were shown a succession of cards, each card either red or blue. Seventy percent of the cards were blue, and

1. See also Schul and Mayo (2003).

30% red; the color sequence was random. The subjects, asked to guess each succeeding card, would guess blue around 70% of the time, and red about 30% of the time—as if they thought they had some way of predicting the random sequence! Even when the subjects were paid a nickel for each correct guess, they still only guessed blue about 76% of the time. Why is this odd? Because you do not need to bet on a guess to test it. You could just pick blue each time, being paid a nickel about 70% of the time, accumulating thirty-five dollars over a thousand trials, while mentally noting your private guesses for any (imaginary) patterns you thought you spotted. If your predictions came out right, *then* you could switch to the newly discovered sequence. There was no need for the subjects to bet on any patterns they thought they saw; they could have simply bet on blue until some hypothesis was *confirmed*. But if human beings reasoned like that, people would not buy lottery tickets, but instead write down predictions in notebooks at home, and begin buying lottery tickets only when their predictions began succeeding.

The mistake revealed by the experiment was not that the subjects looked for patterns in a random-seeming sequence; that is curiosity, an admirable human trait. Dawes (1988) comments on this experiment: "Despite feedback through a thousand trials, subjects cannot bring themselves to believe that the situation is one in which they *cannot* predict." But even if subjects refused to accept unpredictability and continued looking for patterns, they didn't have to *bet* on their guesses. They just needed to make a mental note of the pattern's prediction, then keep betting on blue while waiting for confirmation. My suspicion is that subjects just didn't think of the winning strategy. They didn't realize that their betting pattern did not have to *resemble* the observed sequence of cards. On *each* round, blue is the *most likely* next card. The best financial strategy is not betting a mostly-blue pattern resembling the mostly-blue sequence, but betting all blue, to win as many nickels as possible. If 70% of the time you predict blue and 30% of the time you predict red, and the cards do not correlate with your guesses, you shall predict correctly $0.7 \times 0.7 + 0.3 \times 0.3 = 58\%$ of the time. If 100% of the time you predict blue, you'll get a nickel 70% of the time.

Under conditions of uncertainty, your optimal *betting pattern* doesn't resemble a *typical sequence* of cards. Similarly, I wonder how many betters on horse races realize that you don't win by betting on the horse you think will win the race, but by betting on horses whose payoffs exceed what you think are the odds. But then, statistical thinkers that sophisticated would probably not bet on horse races.

A *proper scoring rule* (another standard math term) is a rule for scoring bets so that you maximize your expected payoff by betting play money that exactly equals the chance of that color flashing. We want a scoring rule so that if the lights actually flash at the frequencies 30% blue, 20% green, and 50% red, you can maximize your average payoff *only* by betting 30 cents on blue, 20 cents on green, and 50 cents on red. A proper

scoring rule is one that forces your optimal bet to exactly report your estimate of the probabilities. (This is also sometimes known as a *strictly proper scoring rule*.) As we've seen, not all scoring rules have this property; and if you invent a plausible-sounding scoring rule at random, it probably *won't* have the property.

One rule with this proper property is to pay a dollar minus the squared error of the bet, rather than the bet itself—if you bet 30 cents on the winning light, your error would be 70 cents, your squared error would be 49 cents, and a dollar minus your squared error would be 51 cents.[2] (Presumably your play money is denominated in the square root of cents, so that the squared error is a monetary sum.)

We shall *not* use the squared-error rule. Ordinary statisticians take the squared error of everything in sight, but not Bayesian statisticians.

We add a new requirement: we require, not only a proper scoring rule, but that our proper scoring rule gives us the same answer whether we apply it to rounds individually or combined. This is what Bayesians do instead of taking the squared error of things; we require invariances.

Suppose I press the button twice in a row. There are nine possible outcomes: green-green, green-blue, green-red, blue-green, blue-blue, blue-red, red-green, red-blue, and red-red. Suppose that green wins, and then blue wins. The experimenter would assign the first score based on our probability assignments for $p(\text{green}_1)$ and the second score based on $p(\text{blue}_2|\text{green}_1)$.[3] We would make two predictions, and get two scores. Our first prediction was the probability we assigned to the color that won on the first round, green. Our second prediction was our probability that blue would win on the second round, *given* that green won on the first round. Why do we need to write $p(\text{blue}_2|\text{green}_1)$ instead of just $p(\text{blue}_2)$? Because you might have a hypothesis about the flashing light that says "blue never follows green," or "blue always follows green" or "blue follows green with 70% probability." If this is so, then after seeing green on the first round, you might want to revise your prediction—change your bets—for the second round. You can always revise your predictions right up to the moment the experimenter presses the button, using every scrap of information; but after the light flashes it is too late to change your bet.

Suppose the actual outcome is green_1 followed by blue_2. We require this invariance: I must get the same total score, regardless of whether:

2. Readers with calculus may verify that in the simpler case of a light that has only two colors, with p being the bet on the first color and f the frequency of the first color, the expected payoff $f \times (1 - (1 - p)^2) + (1 - f) \times (1 - p^2)$, with p variable and f constant, has its global maximum when we set $p = f$.

3. Don't remember how to read $p(A|B)$? See *An Intuitive Explanation of Bayesian Reasoning* (Yudkowsky 2003).

- I am scored twice, first on my prediction for $p(\text{green}_1)$, and second on my prediction for $p(\text{blue}_2|\text{green}_1)$.
- I am scored once for my joint prediction $p(\text{blue}_2 \,\&\, \text{green}_1)$.

Suppose I assign a 60% probability to green_1, and then the green light flashes. I must now produce probabilities for the colors on the second round. I assess the possibility blue_2, and allocate it 25% of my probability mass. Lo and behold, on the second round the light flashes blue. So on the first round my bet on the winning color was 60%, and on the second round my bet on the winning color was 25%. But I might also, at the start of the experiment and after assigning $p(\text{green}_1)$, imagine that the light first flashes green, imagine updating my theories based on that information, and then say what confidence I will give to blue on the next round if the first round is green. That is, I generate the probabilities $p(\text{green}_1)$ and $p(\text{blue}_2|\text{green}_1)$. By multiplying these two probabilities together we would get the joint probability, $p(\text{green}_1 \,\&\, \text{blue}_2) = 15\%$.

A double experiment has nine possible outcomes. If I generate nine probabilities for $p(\text{green}_1 \,\&\, \text{green}_2)$, $p(\text{green}_1 \,\&\, \text{blue}_2)$, . . . , $p(\text{red}_1 \,\&\, \text{blue}_2)$, $p(\text{red}_1 \,\&\, \text{red}_2)$, the probability mass must sum to no more than one. I am giving predictions for nine mutually exclusive possibilities of a "double experiment."

We require a scoring rule (and maybe it won't look like anything an ordinary bookie would ever use) such that my score doesn't change regardless of whether we consider the double result as two predictions or one prediction. I can treat the sequence of two results as a single experiment, "press the button twice," and be scored on my prediction for $p(\text{blue}_2 \,\&\, \text{green}_1) = 15\%$. Or I can be scored once for my first prediction $p(\text{green}_1) = 60\%$, then again on my prediction $p(\text{blue}_2|\text{green}_1) = 25\%$. We require the same *total* score in either case, so that it doesn't matter how we slice up the experiments and the predictions—the *total* score is always exactly the same. This is our invariance.

We have just required:

$$\text{Score}[p(\text{green}_1 \,\&\, \text{blue}_2)] = \text{Score}[p(\text{green}_1)] + \text{Score}[p(\text{blue}_2|\text{green}_1)]$$

And we already know:

$$p(\text{green}_1 \,\&\, \text{blue}_2) = p(\text{green}_1) \times p(\text{blue}_2|\text{green}_1)$$

The only possible scoring rule is:

$$\text{Score}\, p = \log p$$

The new scoring rule is that your score is the *logarithm* of the probability you assigned to the winner.

The base of the logarithm is arbitrary—whether we use the logarithm base ten or the logarithm base two, the scoring rule has the desired invariance. But we must choose

some actual base. A mathematician would choose base e; an engineer would choose base ten; a computer scientist would choose base two. If we use base ten, we can convert to *decibels*, as in the *Intuitive Explanation*; but sometimes bits are easier to manipulate.

The logarithm scoring rule is proper—it has its expected maximum when we say our exact expectations; it rewards honesty. If we think the blue light has a 60% probability of flashing, and we calculate our expected payoff for different betting schemas, we find that we maximize our expected payoff by telling the experimenter "60%." (Readers with calculus can verify this.) The scoring rule also gives an invariant total, regardless of whether pressing the button twice counts as "one experiment" or "two experiments." However, payoffs are now all *negative*, since we are taking the logarithm of the probability and the probability is between zero and one. The logarithm base ten of 0.1 is -1; the logarithm base ten of 0.01 is -2. That's okay. We accepted that the scoring rule might not look like anything a real bookie would ever use. If you like, you can imagine that the experimenter has a pile of money, and at the end of the experiment he awards you some amount minus your large negative score. (Er, the amount plus your negative score.) Maybe the experimenter has a hundred dollars, and at the end of a hundred rounds you accumulated a score of -48, so you get \$52 dollars.

A score of -48 in what base? We can eliminate the ambiguity in the score by specifying units. Ten decibels equals a factor of 10; negative ten decibels equals a factor of 1/10. Assigning a probability of 0.01 to the actual outcome would score -20 decibels. A probability of 0.03 would score -15 decibels. Sometimes we may use bits: 1 bit is a factor of 2, -1 bit is a factor of 1/2. A probability of 0.25 would score -2 bits; a probability of 0.03 would score around -5 bits.

If you arrive at a probability assessment p for each color, with $p(\text{red})$, $p(\text{blue})$, $p(\text{green})$, then your *expected score* is:

$$\text{Score}\, p = \log p$$
$$\text{Expectation}(\text{Score}) = \text{Sum}(p \times \log p) \text{ for all outcomes } p.$$

Suppose you had probabilities of 25% red, 50% blue, and 25% green. Let's think in base 2 for a moment, to make things simpler. Your expected score is:

$$\begin{aligned}
\text{Score}(\text{red}) &= -2 \text{ bits, flashes } 25\% \text{ of the time,} \\
\text{Score}(\text{blue}) &= -1 \text{ bit, flashes } 50\% \text{ of the time,} \\
\text{Score}(\text{green}) &= -2 \text{ bits, flashes } 25\% \text{ of the time,} \\
\text{Expectation}(\text{Score}) &= -1.5 \text{ bits.}
\end{aligned}$$

* * *

Contrast our Bayesian scoring rule with the ordinary or colloquial way of speaking about degrees of belief, where someone might casually say, "I'm 98% certain that canola oil

contains more omega-3 fats than olive oil." What they really mean by this is that they feel 98% certain—there's something like a little progress bar that measures the strength of the emotion of certainty, and this progress bar is 98% full. And the emotional progress bar probably wouldn't be exactly 98% full, if we had some way to measure. The word "98%" is just a colloquial way of saying: "I'm almost but not entirely certain." It doesn't mean that you could get the highest expected payoff by betting exactly 98 cents of play money on that outcome. You should only assign a *calibrated confidence* of 98% if you're confident enough that you think you could answer a hundred similar questions, of equal difficulty, one after the other, each independent from the others, and be wrong, on average, about twice. We'll keep track of how often you're right, over time, and if it turns out that when you say "90% sure" you're right about seven times out of ten, then we'll say you're *poorly calibrated*.

Remember Spock from Star Trek? Spock often says something along the lines of, "Captain, if you steer the Enterprise directly into a black hole, our probability of survival is only 2.837%." Yet nine times out of ten the Enterprise is not destroyed. What kind of tragic fool gives a figure with four significant digits of precision that is wrong by two orders of magnitude?

The people who write this stuff have no idea what scientists mean by "probability." They suppose that a probability of 99.9% is something like feeling really sure. They suppose that Spock's statement expresses the *challenge* of successfully steering the Enterprise through a black hole, like a video game rated five stars for difficulty. What we mean by "probability" is that if you utter the words "two percent probability" on fifty independent occasions, it better not happen more than once.

If you say "98% probable" a thousand times, and you are surprised only five times, we still ding you for poor calibration. You're allocating too much probability mass to the possibility that you're wrong. You should say "99.5% probable" to maximize your score. The scoring rule rewards *accurate* calibration, encouraging neither humility nor arrogance.

At this point it may occur to some readers that there's an obvious way to achieve perfect calibration—just flip a coin for every yes-or-no question, and assign your answer a confidence of 50%. You say 50% and you're right half the time. Isn't that perfect calibration? Yes. But calibration is only one component of our Bayesian score; the other component is *discrimination*.

Suppose I ask you ten yes-or-no questions. You know absolutely nothing about the subject, so on each question you divide your probability mass fifty-fifty between "Yes" and "No." Congratulations, you're perfectly calibrated—answers for which you said "50% probability" were true exactly half the time. This is true regardless of the sequence of correct answers or how many answers were Yes. In ten experiments you said "50%" on

twenty occasions—you said "50%" to Yes_1–No_1, Yes_2–No_2, Yes_3–No_3, . . . On ten of those occasions the answer was correct, the occasions: Yes_1, No_2, No_3, . . . And on ten of those occasions the answer was incorrect: No_1, Yes_2, Yes_3, . . .

Now I give my own answers, putting more effort into it, trying to discriminate whether Yes or No is the correct answer. I assign 90% confidence to each of my favored answers, and my favored answer is wrong twice. I'm more poorly calibrated than you. I said "90%" on ten occasions and I was wrong two times. The next time someone listens to me, they may mentally translate "90%" into 80%, knowing that when I'm 90% sure I'm right about 80% of the time. But the probability you assigned to the final outcome is 1/2 to the tenth power, which is 0.001 or 1/1024. The probability I assigned to the final outcome is 90% to the eighth power times 10% to the second power, $0.9^8 \times 0.1^2$, which works out to 0.004 or 0.4%. Your calibration is perfect and mine isn't, but my better *discrimination* between right and wrong answers more than makes up for it. My final score is higher—I assigned a greater joint probability to the final outcome of the entire experiment. If I'd been less overconfident and better calibrated, the probability I assigned to the final outcome would have been $0.8^8 \times 0.2^2$, which works out to 0.006 or 6%.

Is it possible to do even better? Sure. You could have guessed every single answer correctly, and assigned a probability of 99% to each of your answers. Then the probability you assigned to the entire experimental outcome would be $0.99^{10} \approx 90\%$.

Your *score* would be $\log 90\%$, -0.45 decibels or -0.15 bits. We need to take the logarithm so that if I try to maximize my *expected score*, $\text{Sum}(p \times \log p)$, I have no motive to cheat. Without the logarithm rule, I would maximize my expected score by assigning all my probability mass to the most probable outcome. Also, without the logarithm rule, my total score would be different depending on whether we counted several rounds as several experiments or as one experiment.

A simple transform can fix poor calibration by decreasing discrimination. If you are in the habit of saying "million-to-one" on 90 correct and 10 incorrect answers for each hundred questions, we can perfect your calibration by replacing "million-to-one" with "nine-to-one." In contrast, there's no easy way to increase (successful) discrimination. If you habitually say "nine-to-one" on 90 correct answers for each hundred questions, I can easily increase your *claimed* discrimination by replacing "nine-to-one" with "million-to-one." But no simple transform can increase your *actual* discrimination such that your reply distinguishes 95 correct answers and 5 incorrect answers. From Yates et al. (2002): "Whereas good calibration often can be achieved by simple mathematical transformations (e.g., adding a constant to every probability judgment), good discrimination demands access to solid, predictive evidence and skill at exploiting that evidence, which are difficult to find in any real-life, practical situation." If you lack the ability to distinguish

truth from falsehood, you can achieve perfect calibration by confessing your ignorance; but confessing ignorance will not, of itself, distinguish truth from falsehood.

We thus dispose of another false stereotype of rationality, that rationality consists of being humble and modest and confessing helplessness in the face of the unknown. That's just the cheater's way out, assigning a 50% probability to all yes-or-no questions. Our scoring rule encourages you to do better if you can. If you are ignorant, confess your ignorance; if you are confident, confess your confidence. We penalize you for being confident and wrong, but we also reward you for being confident and right. That is the virtue of a proper scoring rule.

* * *

Suppose I flip a coin twenty times. If I believe the coin is fair, the best prediction I can make is to predict an even chance of heads or tails on each flip. If I believe the coin is fair, I assign the same probability to every possible sequence of twenty coinflips. There are roughly a million (1,048,576) possible sequences of twenty coinflips, and I have only 1.0 of probability mass to play with. So I assign to each *individual* possible sequence a probability of $(1/2)^{20}$—odds of about a million to one; -20 bits or -60 decibels.

I made an experimental prediction and got a score of -60 decibels! Doesn't this falsify the hypothesis? Intuitively, no. We do not flip a coin twenty times and see a random-looking result, then reel back and say, why, the odds of that are a million to one. But the odds *are* a million to one against seeing that exact sequence, as I would discover if I naively predicted the exact same outcome for the *next* sequence of twenty coinflips. It's okay to have theories that assign tiny probabilities to outcomes, so long as no other theory does better. But if someone used an alternate hypothesis to write down the exact sequence in a sealed envelope in advance, and she assigned a probability of 99%, I would suspect the fairness of the coin. Provided that she only sealed *one* envelope, and not a million.

That tells us *what* we ought common-sensically to answer, but it doesn't say *how* the common-sense answer arises from the math. To say *why* the common sense is correct, we need to integrate all that has been said so far into the framework of Bayesian revision of belief. When we're done, we'll have a technical understanding of the difference between a verbal understanding and a technical understanding.

* * *

Imagine an experiment which produces an integer result between zero and 99. For example, the experiment might be a particle counter that tells us how many particles have passed through in a minute. Or the experiment might be to visit the supermarket on Wednesday, check the price of a 10 oz bag of crushed walnuts, and write down the last two digits of the price.

We are testing several different hypotheses that try to predict the experimental result. Each hypothesis produces a probability distribution over all possible results; in this case, the integers between zero and 99. The possibilities are mutually exclusive, so the probability mass in the distribution must sum to one (or less); we cannot predict a 90% probability of seeing 42 and also a 90% probability of seeing 43.

Suppose there is a precise hypothesis, which predicts a 90% chance of seeing the result 51. (I.e., the hypothesis is that the supermarket usually prices walnuts with a price of "X dollars and 51 cents.") The precise theory has staked 90% of its probability mass on the outcome 51. This leaves 10% probability mass remaining to spread over 99 other possible outcomes—all the numbers between zero and 99 *except* 51. The theory makes no further specification, so we spread the remaining 10% probability mass evenly over 99 possibilities, assigning a probability of 1/990 to each non-51 result. For ease of writing, we'll approximate 1/990 as 0.1%.

This probability distribution is analogous to the *likelihood* or *conditional probability* of the result given the hypothesis. Let us call it the *likelihood distribution* for the hypothesis, our chance of seeing each specified outcome *if* the hypothesis is true. The likelihood distribution for a hypothesis H is a function composed of all the conditional probabilities for $p(0|H) = 0.001$, $p(1|H) = 0.001$, . . . , $p(51|H) = 0.9$, . . . , $p(99|H) = 0.001$. The probability mass contained in the likelihood distribution must sum to one. It is a general rule that there is no way we can have a 90% chance of seeing 51 and also a 90% chance of seeing 52. Therefore, if we first assume the hypothesis H is true, there is still no way we can have a 90% chance of seeing 51 and also a 90% chance of seeing 52.

The precise theory predicts a 90% probability of seeing 51. Let there be also a vague theory, which predicts "a 90% probability of seeing a number in the fifties."

Seeing the result 51, we do not say the outcome confirms both theories equally. Both theories made predictions, and both assigned probabilities of 90%, and the result 51 confirms both predictions. But the precise theory has an advantage because it concentrates its probability mass into a sharper point. If the vague theory makes no further specification, we count "a 90% probability of seeing a number in the fifties" as a 9% probability of seeing each number between 50 and 59.

Suppose we started with even odds in favor of the precise theory and the vague theory—odds of 1:1, or 50% probability for either hypothesis being true. After seeing the result 51, what are the posterior odds of the precise theory being true? (If you don't remember how to work this problem, return to *An Intuitive Explanation of Bayesian Reasoning*.) The predictions of the two theories are analogous to their likelihood assignments—the conditional probability of seeing the result, given that the theory is true. What is the likelihood ratio between the two theories? The first theory allocated 90% probability mass to the *exact* outcome. The vague theory allocated 9% probability

mass to the exact outcome. The likelihood ratio is 10:1. So if we started with even 1:1 odds, the posterior odds are 10:1 in favor of the precise theory. The differential pressure of the two conditional probabilities pushed our prior confidence of 50% to a posterior confidence of about 91% that the precise theory is correct. *Assuming* that these are the only hypotheses being tested, that this is the only evidence under consideration, and so on.

Why did the vague theory lose when both theories fit the evidence? The vague theory is timid; it makes a broad prediction, hedges its bets, allows many possibilities that would falsify the precise theory. This is not the virtue of a scientific theory. Philosophers of science tell us that theories should be bold, and subject themselves willingly to falsification if their prediction fails (Popper 1959). Now we see why. The precise theory concentrates its probability mass into a sharper point and thereby leaves itself vulnerable to falsification if the real outcome hits elsewhere; but if the predicted outcome is correct, precision has a tremendous likelihood advantage over vagueness.

The laws of probability theory provide no way to cheat, to make a vague hypothesis such that any result between 50 and 59 counts for as much favorable confirmation as the precise theory receives, for that would require probability mass summing to 900%. There is no way to cheat, providing you record your prediction *in advance*, so you cannot claim afterward that your theory assigns a probability of 90% to whichever result arrived. Humans are very fond of making their predictions afterward, so the social process of science requires an advance prediction before we say that a result confirms a theory. But how humans may move in harmony with the way of Bayes, and so wield the power, is a separate issue from whether the math works. When we're doing the math, we just take for granted that likelihood density functions are fixed properties of a hypothesis and the probability mass sums to 1 and you'd never dream of doing it any other way.

You may want to take a moment to visualize that, *if* we define probability in terms of calibration, Bayes' Theorem relates the calibrations. Suppose I guess that Theory 1 is 50% likely to be true, and I guess that Theory 2 is 50% likely to be true. Suppose I am well-calibrated; when I utter the words "fifty percent," the event happens about half the time. And then I see a result R which would happen around nine-tenths of the time given Theory 1, and around nine-hundredths of the time given Theory 2, and I know this is so, and I apply Bayesian reasoning. If I was perfectly calibrated initially (despite the poor discrimination of saying 50/50), I will still be perfectly calibrated (and better discriminated) after I say that my confidence in Theory 1 is now 91%. If I repeated this kind of situation many times, I would be right around ten-elevenths of the time when I said "91%." If I reason using Bayesian rules, and I start from well-calibrated priors, then my conclusions will also be well-calibrated. This only holds true if we define probability in terms of calibration! If "90% sure" is instead interpreted as, say, the strength of the

emotion of surety, there is no reason to expect the posterior emotion to stand in an exact Bayesian relation to the prior emotion.

Let the prior odds be ten to one in favor of the vague theory. Why? Suppose our way of describing hypotheses allows us to either specify a precise number, or to just specify a first-digit; we can say "51," "63," "72," or "in the fifties/sixties/seventies." Suppose we think that the real answer is about equally liable to be an answer of the first kind or the second. However, given the problem, there are a hundred possible hypotheses of the first kind, and only ten hypotheses of the second kind. So if we think that either *class* of hypotheses has about an equal prior chance of being correct, we have to spread out the prior probability mass over ten times as many precise theories as vague theories. The precise theory that predicts exactly 51 would thus have one-tenth as much prior probability mass as the vague theory that predicts a number in the fifties. After seeing 51, the odds would go from 1:10 in favor of the vague theory to 1:1, even odds for the precise theory and the vague theory.

If you look at this carefully, it's exactly what common sense would expect. You start out uncertain of whether a phenomenon is the kind of phenomenon that produces exactly the same result every time, or if it's the kind of phenomenon that produces a result in the *X*ties every time. (Maybe the phenomenon is a price range at the supermarket, if you need some reason to suppose that 50–59 is an acceptable range but 49–58 isn't.) You take a single measurement and the answer is 51. Well, that could be because the phenomenon is exactly 51, or because it's in the fifties. So the remaining precise theory has the same odds as the remaining vague theory, which requires that the vague theory must have started out ten times as probable as that precise theory, since the precise theory has a sharper fit to the evidence.

If we just see one number, like 51, it doesn't change the prior probability that the phenomenon itself was "precise" or "vague." But, in effect, it concentrates all the probability mass of those two *classes* of hypothesis into a single surviving hypothesis of each class.

Of course, it is a severe error to say that a *phenomenon* is precise or vague, a case of what Jaynes (2003) calls the Mind Projection Fallacy. Precision or vagueness is a property of maps, not territories. Rather we should ask if the price in the supermarket stays constant or shifts about. A hypothesis of the "vague" sort is a good description of a price that shifts about. A precise map will suit a constant territory.

Another example: You flip a coin ten times and see the sequence HHTTH:TTTTH. Maybe you started out thinking there was a 1% chance this coin was fixed. Doesn't the hypothesis "This coin is fixed to produce HHTTH:TTTTH" assign a thousand times the likelihood mass to the observed outcome, compared to the fair coin hypothesis? Yes. Don't the posterior odds that the coin is fixed go to 10:1? No. The 1% prior

probability that "the coin is fixed" has to cover every possible kind of fixed coin—a coin fixed to produce HHTTH:TTTTH, a coin fixed to produce TTHHT:HHHHT, etc. The prior probability the coin is fixed to produce HHTTH:TTTTH is not 1%, but a thousandth of one percent. Afterward, the posterior probability the coin is fixed to produce HHTTH:TTTTH is one percent. Which is to say: You thought the coin was probably fair but had a one percent chance of being fixed to some random sequence; you flipped the coin; the coin produced a random-looking sequence; and that doesn't tell you anything about whether the coin is fair or fixed. It does tell you, if the coin is fixed, *which* sequence it is fixed to.

This parable helps illustrate why Bayesians *must* think about prior probabilities. There is a branch of statistics, sometimes called "orthodox" or "classical" statistics, which insists on paying attention only to likelihoods. But if you only pay attention to likelihoods, then eventually some fixed-coin hypothesis will always defeat the fair coin hypothesis, a phenomenon known as "overfitting" the theory to the data. After thirty flips, the *likelihood* is a billion times as great for the fixed-coin hypothesis with that sequence, as for the fair coin hypothesis. Only if the fixed-coin hypothesis (or rather, that specific fixed-coin hypothesis) is a billion times less probable *a priori*, can the fixed-coin hypothesis possibly lose to the fair coin hypothesis.

If you shake the coin to reset it, and start flipping the coin *again*, and the coin produces HHTTH:TTTTH *again*, that is a different matter. That does raise the posterior odds of the fixed-coin hypothesis to 10:1, even if the starting probability was only 1%.

Similarly, if we perform two successive measurements of the particle counter (or the supermarket price on Wednesdays), and *both* measurements return 51, the precise theory wins by odds of 10:1.

So the precise theory wins, but the vague theory would still score better than no theory at all. Consider a third theory, the hypothesis of zero knowledge or *maximum-entropy distribution*, which makes equally probable any result between zero and 99. Suppose we see the result 51. The vague theory produced a better prediction than the maximum-entropy distribution—assigned a greater likelihood to the outcome we observed. The vague theory is, literally, better than nothing. Suppose we started with odds of 1:20 in favor of the hypothesis of complete ignorance. (Why odds of 1:20? There is only one hypothesis of complete ignorance, and moreover, it's a particularly simple and intuitive kind of hypothesis, Occam's Razor.) After seeing the result of 51, predicted at 9% by the vague theory versus 1% by complete ignorance, the posterior odds go to 10:20 or 1:2. If we then see another result of 51, the posterior odds go to 10:2 or 83% probability for the vague theory, assuming there is no more precise theory under consideration.

Yet the timidity of the vague theory—its unwillingness to produce an *exact* prediction and accept falsification on any other result—renders it vulnerable to the bold, precise theory. (Providing, of course, that the bold theory correctly guesses the outcome!) Suppose the prior odds were 1:10:200 for the precise, vague, and ignorant theories—prior probabilities of 0.5%, 4.7%, and 94.8% for the precise, vague and ignorant theories. This figure reflects our prior probability distribution over *classes* of hypotheses, with the probability mass distributed over entire classes as follows: 50% that the phenomenon shifts across all digits, 25% that the phenomenon shifts around within some decimal bracket, and 25% that the phenomenon repeats the same number each time. One hypothesis of complete ignorance, 10 possible hypotheses for a decimal bracket, 100 possible hypotheses for a repeating number. Thus, prior odds of 1:10:200 for the precise hypothesis 51, the vague hypothesis "fifties," and the hypothesis of complete ignorance.

After seeing a result of 51, with assigned probability of 90%, 9%, and 1%, the posterior odds go to 90:90:200 = 9:9:20. After seeing an additional result of 51, the posterior odds go to 810:81:20, or 89%, 9%, and 2%. The precise theory is now favored over the vague theory, which in turn is favored over the ignorant theory.

Now consider a stupid theory, which predicts a 90% probability of seeing a result between zero and nine. The stupid theory assigns a probability of 0.1% to the actual outcome, 51. If the odds were initially 1:10:200:10 for the precise, vague, ignorant, and stupid theories, the posterior odds after seeing 51 once would be 90:90:200:1. The stupid theory has been falsified (posterior probability of 0.2%).

It is possible to have a model so bad that it is worse than nothing, if the model concentrates its probability mass away from the actual outcome, makes confident predictions of wrong answers. Such a hypothesis is so poor that it loses against the hypothesis of complete ignorance. Ignorance is better than anti-knowledge.

> *Side note 1:* In the field of Artificial Intelligence, there is a sometime fad that praises the glory of randomness. Occasionally an AI researcher discovers that if they add noise to one of their algorithms, the algorithm works better. This result is reported with great enthusiasm, followed by much fulsome praise of the creative powers of chaos, unpredictability, spontaneity, ignorance of what your own AI is doing, et cetera. (See Imagination Engines, Inc. [2011] for an example; according to their sales literature they sell wounded and dying neural nets.) But how sad is an algorithm if you can *increase* its performance by injecting entropy into intermediate processing stages? The algorithm must be so deranged that some of its work goes into concentrating probability mass *away* from good solutions. If injecting randomness results in a reliable improvement, then some aspect of the algorithm must do reliably worse than random. Only in AI would people devise algorithms *literally dumber than a*

bag of bricks, boost the results slightly back toward ignorance, and then argue for the healing power of noise.

Side note 2: Robert Pirsig (1974) once said: "The world's stupidest man may say the Sun is shining, but that doesn't make it dark out". It is a classical logical fallacy to say, "Hitler believed in the Pythagorean Theorem. You don't want to agree with Hitler, do you?" Consider that for someone to be reliably wrong on yes-or-no questions—say, to be wrong 90% of the time—that person would need to do all the hard work of discriminating truth from falsehood, just to be wrong so reliably. If someone is wrong on yes-or-no questions 99% of the time, we can get 99% accuracy just by inverting the responses. Anyone that stupid would be smarter than I am.

Suppose that in our experiment we see the results 52, 51, and 58. The precise theory gives this conjunctive event a probability of a thousand to one times 90% times a thousand to one, while the vaguer theory gives this conjunctive event a probability of 9% cubed, which works out to . . . oh . . . um . . . let's see . . . a million to one given the precise theory, versus a thousand to one given the vague theory. Or thereabouts; we are counting rough powers of ten. Versus a million to one given the zero-knowledge distribution that assigns an equal probability to all outcomes. Versus a billion to one given a model worse than nothing, the stupid hypothesis, which claims a 90% probability of seeing a number less than 10. Using these approximate numbers, the vague theory racks up a score of -30 decibels (a probability of 1/1000 for the whole experimental outcome), versus scores of -60 for the precise theory, -60 for the ignorant theory, and -90 for the stupid theory. It is not always true that the highest score wins, because we need to take into account our prior odds of 1:10:200:10, confidences of -23, -13, 0, and -13 decibels. The vague theory still comes in with the highest total score at -43 decibels. (If we ignored our prior probabilities, each new experiment would override the accumulated results of all the previous experiments; we could not accumulate knowledge. Furthermore, the fixed-coin hypothesis would always win.)

As always, we should not be alarmed that even the best theory still has a low score—recall the parable of the fair coin. Theories are approximations. In principle we might be able to predict the exact sequence of coinflips. But it would take better measurement and more computing power than we're willing to expend. Maybe we could achieve 60/40 prediction of coinflips, with a good enough model . . . ? We go with the best approximation we have, and try to achieve good calibration even if the discrimination isn't perfect.

* * *

We've conducted our analysis so far under the rules of Bayesian probability theory, in which there's no way to have more than 100% probability mass, and hence no way to

cheat so that any outcome can count as "confirmation" of your theory. Under Bayesian law, play money may not be counterfeited; you only have so much clay. If you allocate more probability mass in one place, you have to take it from somewhere else; a coin cannot have a 90% chance of turning up heads and a 90% chance of turning up tails.

Unfortunately, human beings are not Bayesians. Human beings bizarrely attempt to *defend* hypotheses, making a deliberate effort to prove them or prevent disproof. This behavior has no analogue in the laws of probability theory or decision theory. In formal probability theory the hypothesis *is*, and the evidence *is*, and either the hypothesis is confirmed or it is not. In formal decision theory, an agent may make an effort to investigate some issue of which the agent is currently uncertain, not knowing whether the evidence shall go one way or the other. In neither case does one ever deliberately try to prove an idea, or try to avoid disproving it. One may *test* ideas of which one is genuinely uncertain, but not have a "preferred" outcome of the investigation. One may not try to prove hypotheses, nor prevent their proof. I cannot properly convey just how ridiculous the notion would be, to a true Bayesian; there are not even words in Bayes-language to describe the mistake . . .

One classic method for preventing a theory from disproof is arguing *post facto* that any observation presented proves the theory. Friedrich Spee ([1631] 2003), hearing the confessions of many condemned witches as a priest, wrote the *Cautio Criminalis* ("prudence in criminal cases") in which he bitingly described the decision tree for condemning accused witches. If the witch had led an evil and improper life, she was guilty; if she had led a good and proper life, this too was a proof, for witches dissemble and try to appear especially virtuous. After the woman was put in prison: if she was afraid, this proved her guilt; if she was not afraid, this proved her guilt, for witches characteristically pretend innocence and wear a bold front. Or on hearing of a denunciation of witchcraft against her, she might seek flight or remain; if she ran, that proved her guilt; if she remained, the devil had detained her so she could not get away. Spee acted as confessor to many witches; he was thus in a position to observe *every* branch of the accusation tree, that no matter *what* the accused witch said or did, it was held a proof against her. In any individual case, you would only hear one branch of the dilemma.

It is for this reason that scientists write down their predictions in advance.

If you've read the *Intuitive Explanation*, you should recall the result I nicknamed the "Law of Conservation of Probability," that for every expectation of evidence there is an equal and opposite expectation of counterevidence. If A is evidence in favor of B, *not-A must* be evidence in favor of *not-B*. The strengths of the evidences may not be equal; rare but strong evidence in one direction may be balanced by common but weak evidence in the other direction. But it is not possible for both A and *not-A* to be evidence in favor of B. That is, it's not possible under the laws of probability theory. Humans often seem

to want to have their cake and eat it too. Whichever result we witness is the one that proves our theory. As Spee ([1631] 2003) put it, "The investigating committee would feel disgraced if it acquitted a woman; once arrested and in chains, she has to be guilty, by fair means or foul."

The way human psychology seems to work is that first we see something happen, and then we try to argue that it matches whatever hypothesis we had in mind beforehand. Rather than conserved probability mass, to distribute over advance *predictions*, we have a feeling of *compatibility*—the degree to which the explanation and the event seem to "fit." "Fit" is not conserved. There is no equivalent of the rule that probability mass must sum to one. A psychoanalyst may explain any possible behavior of a patient by constructing an appropriate structure of "rationalizations" and "defenses"; it fits, therefore it must be true.

Now consider the fable told at the start of this essay—the students seeing a radiator, and a metal plate next to the radiator. The students would never predict in advance that the side of the plate near the radiator would be cooler. Yet, seeing the fact, they managed to make their explanations "fit." They lost their precious chance at bewilderment, to realize that their models did not predict the phenomenon they observed. They sacrificed their ability to be more confused by fiction than by truth. And they did not realize "heat induction, blah blah, therefore the near side is cooler" is a vague and verbal prediction, spread across an enormously wide range of possible values for specific measured temperatures. Applying equations of diffusion and equilibrium would give a *sharp* prediction for possible joint values. It might not specify the *first* values you measured, but when you knew a few values you could generate a sharp prediction for the rest. The score for the entire experimental outcome would be far better than any less precise alternative, especially a vague and verbal prediction.

* * *

You now have a *technical* explanation of the difference between a verbal explanation and a technical explanation. It is a technical explanation because it enables you to calculate *exactly how technical* an explanation is. Vague hypotheses may be so vague that only a superhuman intelligence could calculate exactly how vague. Perhaps a sufficiently huge intelligence could extrapolate every possible experimental result, and extrapolate every possible verdict of the vague guesser for how well the vague hypothesis "fit," and then renormalize the "fit" distribution into a likelihood distribution that summed to one. But in principle one can still calculate exactly how vague is a vague hypothesis. The calculation is just not computationally tractable, the way that calculating airplane trajectories via quantum mechanics is not computationally tractable.

I hold that everyone needs to learn at least one technical subject: physics, computer science, evolutionary biology, Bayesian probability theory, or *something*. Someone with

no technical subjects under their belt has no referent for what it means to "explain" something. They may think "All is Fire" is an explanation. Therefore do I advocate that Bayesian probability theory should be taught in high school. Bayesian probability theory is the sole piece of math I know that is accessible at the high school level, and that permits a *technical* understanding of a subject matter—the dynamics of belief—that is an everyday real-world domain and has emotionally meaningful consequences. Studying Bayesian probability would give students a referent for what it means to "explain" something.

Too many academics think that being "technical" means speaking in dry polysyllabisms. Here's a "technical" explanation of technical explanation:

> The equations of probability theory favor hypotheses that strongly predict the exact observed data. Strong models boldly concentrate their probability density into precise outcomes, making them falsifiable if the data hits elsewhere, and giving them tremendous likelihood advantages over models less bold, less precise. Verbal explanation runs on psychological evaluation of unconserved post facto compatibility instead of conserved ante facto probability density. And verbal explanation does not paint sharply detailed pictures, implying a smooth likelihood distribution in the vicinity of the data.

Is this satisfactory? No. Hear the impressive and weighty sentences, resounding with the dull thud of expertise. See the hapless students, writing those sentences on a sheet of paper. Even after the listeners hear the ritual words, they can perform no calculations. *You* know the math, so the words are meaningful. You can perform the calculations after hearing the impressive words, just as you could have done before. But what of one who did not see any calculations performed? What new skills have they gained from that "technical" lecture, save the ability to recite fascinating words?

"Bayesian" sure is a fascinating word, isn't it? Let's get it out of our systems: Bayes Bayes Bayes Bayes Bayes Bayes Bayes Bayes Bayes . . .

The sacred syllable is meaningless, except insofar as it tells someone to apply math. Therefore the one who hears must already know the math.

Conversely, if you know the math, you can be as silly as you like, and still technical.

We thus dispose of yet another stereotype of rationality, that rationality consists of sere formality and humorless solemnity. What has that to do with the problem of distinguishing truth from falsehood? What has that to do with attaining the map that reflects the territory? A scientist worthy of a lab coat should be able to make original discoveries while wearing a clown suit, or give a lecture in a high squeaky voice from inhaling helium. It is written nowhere in the math of probability theory that one may have no fun. The blade that cuts through to the correct answer has no dignity or silliness

of itself, though it may fit the hand of a silly wielder.

* * *

Our physics uses the same *theory* to describe an airplane, and collisions in a particle accelerator—particles and airplanes both obey special relativity and general relativity and quantum electrodynamics and quantum chromodynamics. But we use entirely different *models* to understand the aerodynamics of a 747 and a collision between gold nuclei. A computer modeling the aerodynamics of the 747 may not contain a single token representing an atom, even though no one denies that the 747 is made of atoms.

A *useful* model isn't just something you know, as you know that the airplane is made of atoms. A useful model is knowledge you can compute in reasonable time to predict real-world events you know how to observe. Physicists use different models to predict airplanes and particle collisions, not because the two events take place in different universes with different laws of physics, but because it would be too expensive to compute the airplane particle by particle.

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. It doesn't change the territory. The precision or vagueness of the map isn't a fact about the territory, it's a fact about the map.

Maybe someone will find that, using a model that violates conservation of momentum just a little, you can compute the aerodynamics of the 747 much more *cheaply* than if you insist that momentum is exactly conserved. So if you've got two computers competing to produce the best prediction, it might be that the best prediction comes from the model that violates conservation of momentum. This doesn't mean that the 747 violates conservation of momentum in real life. Neither model uses individual atoms, but that doesn't imply the 747 is not made of atoms. You would prove the 747 is made of atoms with experimental data that the aerodynamic models couldn't handle; for example, you would train a scanning tunneling microscope on a section of wing and look at the atoms. Similarly, you could use a finer measuring instrument to discriminate between a 747 that *really* disobeyed conservation of momentum like the cheap approximation predicted, versus a 747 that obeyed conservation of momentum like underlying physics predicted. The winning theory is the one that best predicts all the experimental predictions together. Our Bayesian scoring rule gives us a way to combine the results of *all* our experiments, even experiments that use different methods.

Furthermore, the atomic theory allows, embraces, and in some sense mandates the aerodynamic model. By thinking abstractly about the assumptions of atomic theory, we realize that the aerodynamic model ought to be a good (and much cheaper) approximation of the atomic theory, and so the atomic theory supports the aerodynamic model, rather than competing with it. A successful theory can embrace many models for differ-

ent domains, so long as the models are acknowledged as approximations, and in each case the model is compatible with (or ideally mandated by) the underlying theory.

Our *fundamental* physics—quantum mechanics, the standard family of particles, and relativity—is a theory that embraces an *enormous* family of models for macroscopic physical phenomena. There is the physics of liquids, and solids, and gases; yet this does not mean that there are *fundamental* things in the world that have the intrinsic property of liquidity.

> Apparently there is colour, apparently sweetness, apparently bitterness, actually there are only atoms and the void.
>
> —Democritus, 420 BC, (quoted in Robinson and Groves 1998)

* * *

In arguing that a "technical" theory should be defined as a theory that sharply concentrates probability into specific advance predictions, I am setting an extremely high standard of strictness. We have seen that a vague theory *can* be better than nothing. A vague theory can win out over the hypothesis of ignorance, if there are no precise theories to compete against it.

There is an enormous family of models belonging to the central underlying theory of life and biology; the underlying theory that is sometimes called neo-Darwinism, natural selection, or evolution. Some models in evolutionary theory are quantitative. The way in which DNA encodes proteins is redundant; two different DNA sequences can code for exactly the same protein. There are four DNA bases {ATCG} and 64 possible combinations of three DNA bases. But those 64 possible codons describe only 20 amino acids plus a stop code. Genetic drift ought therefore to produce non-functional changes in species genomes, through mutations which by chance become fixed in the gene pool. The accumulation rate of non-functional differences between the genomes of two species with a common ancestor, depends on such parameters as the number of generations elapsed and the intensity of selection at that genetic locus. That's an example of a member of the family of evolutionary models that produces quantitative predictions. There are also disequilibrium allele frequencies under selection, stable equilibria for game-theoretical strategies, sex ratios, etc.

This all comes under the heading of "fascinating words." Unfortunately, there are certain religious factions that spread gross disinformation about evolutionary theory. So I emphasize that many models within evolutionary theory make quantitative predictions that are experimentally confirmed, and that such models are far more than sufficient to demonstrate that, e.g., humans and chimpanzees are related by a common ancestor. If you've been victimized by creationist disinformation—that is, if you've heard any

suggestion that evolutionary theory is controversial or untestable or "just a theory" or non-rigorous or non-technical or in any way not confirmed by an unimaginably huge mound of experimental evidence—I recommend reading the *TalkOrigins FAQ* (TalkOrigins Foundation 2012) and studying evolutionary biology with math.

But imagine going back in time to the nineteenth century, when the theory of natural selection had only just been discovered by Charles Darwin and Alfred Russel Wallace. Imagine evolutionism just after its birth, when the theory had nothing remotely like the modern-day body of quantitative models and great heaping mountains of experimental evidence. There was no way of knowing that humans and chimpanzees would be discovered to have 95% shared genetic material. No one knew that DNA existed. Yet even so, scientists flocked to the new theory of natural selection. And later it turned out that there *was* a precisely copied genetic material with the potential to mutate, that humans and chimps were provably related, etc.

So the very strict, very high standard that I proposed for a "technical" theory is too strict. Historically, it *has* been possible to successfully discriminate true theories from false theories, based on predictions of the sort I called "vague." Vague predictions of, say, 80% confidence, can build up a huge advantage over alternate hypotheses, given enough experiments. Perhaps a theory of this kind, producing predictions that are not precisely detailed but are nonetheless correct, could be called "semitechnical"?

But surely technical theories are more reliable than semitechnical theories? Surely technical theories should take precedence, command greater respect? Surely physics, which produces exceedingly exact predictions, is in some sense better confirmed than evolutionary theory? Not implying that evolutionary theory is wrong, of course; but however vast the mountains of evidence favoring evolution, does not physics go one better through vast mountains of *precise* experimental confirmation? Observations of neutron stars confirm the predictions of General Relativity to within one part in a hundred trillion (10^{14}). What does evolutionary theory have to match that?

Someone—I think either Roger Penrose or Richard Dawkins—said once that measured by the simplicity of the theory and the amount of complexity it explained, Darwin had the single greatest idea in the history of time.

Once there was a conflict between nineteenth century physics and nineteenth century evolutionism. According to the best physical models then in use, the Sun could not have been burning very long. Three thousand years on chemical energy, or 40 million years on gravitational energy. There was no energy source known to nineteenth century physics that would permit longer burning. Nineteenth century physics was not *quite* as powerful as modern physics—it did not have predictions accurate to within one part in 10^{14}. But nineteenth century physics still had the mathematical character of modern physics; a discipline whose models produced detailed, precise, quantitative predictions.

Nineteenth century evolutionary theory was wholly semitechnical, without a scrap of quantitative modeling. Not even Mendel's experiments with peas were then known. And yet it did seem likely that evolution would require longer than a paltry 40 million years in which to operate—hundreds of millions, even billions of years. The antiquity of the Earth was a vague and semitechnical prediction, of a vague and semitechnical theory. In contrast, the nineteenth century physicists had a precise and quantitative model, which through formal calculation produced the precise and quantitative dictum that the Sun simply could not have burned that long.

> The limitations of geological periods, imposed by physical science, cannot, of course, disprove the hypothesis of transmutation of species; but it does seem sufficient to disprove the doctrine that transmutation has taken place through "descent with modification by natural selection."
>
> —Lord Kelvin, (quoted in Zapato 2008)

History records who won.

The moral? If you can give 80% confident advance predictions on yes-or-no questions, it may be a "vague" theory, it may be wrong one time out of five, but you can still build up a heck of a huge scoring lead over the hypothesis of ignorance. Enough to confirm a theory, if there are no better competitors. Reality is consistent; every *correct* theory about the universe is compatible with every other correct theory. Imperfect maps can conflict, but there is only one territory. Nineteenth century evolutionism might have been a semitechnical discipline, but it was still correct (as we now know) and by far the best explanation (even in that day). Any conflict between evolutionism and another well-confirmed theory had to reflect some kind of anomaly, a mistake in the assertion that the two theories were incompatible. Nineteenth century physics couldn't model the dynamics of the Sun—they didn't know about nuclear reactions. They could not show that their understanding of the Sun was correct *in technical detail*, nor calculate from a *confirmed* model of the Sun to determine how long the Sun had existed. So in retrospect, we can say something like: "There was room for the possibility that nineteenth century physics just didn't understand the Sun."

But that is hindsight. The real lesson is that, even though nineteenth century physics was both precise and quantitative, it didn't automatically dominate the semitechnical theory of nineteenth century evolutionism. The theories were *both* well-supported. They were *both* correct in the domains over which they were generalized. The apparent conflict between them was an anomaly, and the anomaly turned out to stem from the incompleteness and incorrect application of nineteenth century physics, not the incompleteness and incorrect application of nineteenth century evolutionism. But it would be futile to compare the mountain of evidence supporting the one theory, versus the mountain

of evidence supporting the other. Even in that day, both mountains were too large to suppose that either theory was simply mistaken. Mountains of evidence that large cannot be set to compete, as if one falsifies the other. You must be applying one theory incorrectly, or applying a model outside the domain it predicts well.

So you shouldn't *necessarily* sneer at a theory just because it's semitechnical. Semitechnical theories can build up high enough scores, compared to every available alternative, that you know the theory is at least approximately correct. Someday the semitechnical theory may be replaced or even falsified by a more precise competitor, but that's true even of technical theories. Think of how Einstein's General Relativity devoured Newton's theory of gravitation.

But the correctness of a semitechnical theory—a theory that currently has no precise, computationally tractable models testable by feasible experiments—can be a lot less cut-and-dried than the correctness of a technical theory. It takes skill, patience, and examination to distinguish good semitechnical theories from theories that are just plain confused. This is not something that humans do well by instinct, which is why we have Science.

People eagerly jump the gun and seize on any available reason to reject a disliked theory. That is why I gave the example of nineteenth century evolutionism, to show why one should not be too quick to reject a "non-technical" theory out of hand. By the moral customs of science, nineteenth century evolutionism was guilty of more than one sin. Nineteenth century evolutionism made no quantitative predictions. It was not readily subject to falsification. It was largely an explanation of what had already been seen. It lacked an underlying mechanism, as no one then knew about DNA. It even contradicted the nineteenth century laws of physics. Yet natural selection was such an *amazingly good* post facto explanation that people flocked to it, and they turned out to be right. Science, as a human endeavor, requires advance prediction. Probability theory, as math, does not distinguish between post facto and advance prediction, because probability theory assumes that probability distributions are fixed properties of a hypothesis.

The rule about advance prediction is a rule of the social process of science—a moral custom and not a theorem. The moral custom exists to prevent human beings from making human mistakes that are hard to even describe in the language of probability theory, like tinkering after the fact with what you claim your hypothesis predicts. People concluded that nineteenth century evolutionism was an excellent explanation, even if it was post facto. That reasoning *was correct as probability theory*, which is why it *worked* despite all scientific sins. Probability theory is math. The social process of science is a set of legal conventions to keep people from cheating on the math.

Yet it is also true that, compared to a *modern-day* evolutionary theorist, evolutionary theorists of the late nineteenth and early twentieth century often went sadly astray.

Darwin, who was bright enough to invent the theory, got an amazing amount right. But Darwin's successors, who were only bright enough to accept the theory, misunderstood evolution frequently and seriously. The usual process of science was then required to correct their mistakes. It is incredible how few errors of reasoning Darwin (1859, 1874) made in *The Origin of Species* and *The Descent of Man*, compared to they who followed.

That is also a hazard of a semitechnical theory. Even after the flash of genius insight is confirmed, merely average scientists may fail to apply the insights properly in the absence of formal models. As late as the 1960s biologists spoke of evolution working "for the good of the species," or suggested that individuals would restrain their reproduction to prevent species overpopulation of a habitat. The best evolutionary theorists knew better, but average theorists did not (Williams 1966).

So it is *far* better to have a technical theory than a semitechnical theory. Unfortunately, Nature is not always so kind as to render Herself describable by neat, formal, *computationally tractable* models, nor does She always provide Her students with measuring instruments that can directly probe Her phenomena. Sometimes it is only a matter of time. Nineteenth century evolutionism was semitechnical, but later came the math of population genetics, and eventually DNA sequencing. But Nature will not always give you a phenomenon that you can describe with technical models fifteen seconds after you have the basic insight.

Yet the cutting edge of science, the *controversy*, is most often about a semitechnical theory, or nonsense posing as a semitechnical theory. By the time a theory achieves technical status, it is usually no longer controversial (among scientists). So the question of how to distinguish good semitechnical theories from nonsense is very important to scientists, and it is not as easy as dismissing out of hand any theory that is not technical. To the end of distinguishing truth from falsehood exists the entire discipline of rationality. The art is not reducible to a checklist, or at least, no checklist that an average scientist can apply reliably after an hour of training. If it was that simple we wouldn't need science.

* * *

Why do you care about scientific controversies?

No, seriously, why do you care about scientific *controversies*?

The media thinks that only the cutting edge of science, the very latest controversies, are worth reporting on. How often do you see headlines like "General Relativity still governing planetary orbits" or "Phlogiston theory remains false"? By the time anything is solid science, it is no longer a breaking headline. "Newsworthy" science is 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 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 the problem controversial and attracts all the media attention. So the reporters show up, and hear the scientists speak fascinating words. The reporters are told that "particles" are "waves," but there is no understanding of math for the words to invoke. What the physicist means by "wave" is not what the reporters hear, even if the physicist's math applies also to the structure of water as it crashes on the shore.

And then the reporters write stories, which are not worth the lives of the dead trees on which they are printed.

But what does it matter to you? Why should you pay attention to scientific *controversies*? Why graze upon such sparse and rotten feed as the media offers, when there are so many solid meals to be found in textbooks? Nothing you'll read as breaking news will ever hold a candle to the sheer beauty of settled science. Textbook science has carefully phrased explanations for new students, math derived step by step, plenty of experiments as illustration, and test problems.

And textbook science is beautiful! Textbook science is *comprehensible*, unlike mere fascinating words that can never be truly beautiful. Elementary science textbooks describe *simple* theories, and simplicity is the core of scientific beauty. Fascinating words have no power, nor yet any meaning, without the math. The fascinating words are not knowledge but the illusion of knowledge, which is why it brings so little satisfaction to know that "gravity results from the curvature of spacetime." Science is not in the fascinating words, though it's all the media will ever give you.

Is there ever justification for following a scientific controversy, while there remains any basic science you do not yet know? Yes. You could be an expert in that field, in which case that scientific controversy is your proper meat. Or the scientific controversy might be something you need to know *now*, because it affects your life. Maybe it's the nineteenth century, and you're gazing lustfully at a member of the appropriate sex wearing a nineteenth century bathing suit, and you need to know whether your sexual desire comes from a psychology constructed by natural selection, or is a temptation placed in you by the Devil to lure you into hellfire.

It is not wholly impossible that we shall happen upon a scientific controversy that affects us, and find that we have a burning and urgent need for the correct answer. I shall therefore discuss some of the warning signs that historically distinguished vague hypotheses that later turned out to be unscientific gibberish, from vague hypotheses that later graduated to confirmed theories. Just remember the historical lesson of nineteenth century evolutionism, and resist the temptation to fail every theory that misses a single item on your checklist. It is not my intention to give people another excuse to dismiss good science that discomforts them. If you apply stricter criteria to theories you dislike than theories you like (or vice versa!), then every additional nit you learn how to pick, every new logical flaw you learn how to detect, makes you that much stupider.

Intelligence, to be useful, must be used for something other than defeating itself.

* * *

One of the classic signs of a poor hypothesis is that it must expend great effort in avoiding falsification—elaborating reasons why the hypothesis is compatible with the phenomenon, even though the phenomenon didn't behave as expected.

Sagan (1995) gives the example of someone who claims that a dragon lives in their garage. Fascinated by this controversial question, we ignore all the textbooks providing total solutions to ancient mysteries on which alchemists spent their lives in vain . . . but never mind. We show up at the garage, look inside, and see: Nothing.

Ah, says the claimant, that's because it's an *invisible* dragon.

Now as Sagan says, this is an odd claim, but it doesn't mean we can never know if the dragon is there. Maybe we hear heavy breathing, and discover that carbon dioxide and heat appears in the garage's air. Clawed footprints stamp through the dust. Occasionally a great gout of fire bursts out from no visible source. If so, we conclude that the garage contains an invisible dragon, and the reporters depart, satisfied that the controversy is over. Once something is a fact, it's no longer exciting; it's no fun believing in things that any old fool can see are true. If the dragon were really there, it would be no more fun to believe in the dragon than to believe in zebras.

But now suppose instead that we bring up our measuring instruments to see if carbon dioxide is accumulating in the garage's air, and the claimant at once says: "No, no, it's an invisible non-breathing dragon!" Okay. We begin to examine the dirt, and the claimant says: "No, it's a flying invisible non-breathing dragon, so it won't leave footprints." We start to unload audio equipment, and the claimant says it's an inaudible dragon. We bring in a bag of flour, to throw into the air to outline the dragon's form, and the claimant quickly says that this dragon is permeable to flour.

Carl Sagan originally drew the lesson that poor hypotheses need to do fast footwork to avoid falsification—to maintain an appearance of "fit."

I would point out that the claimant obviously has a good model of the situation *somewhere* in his head, because he can predict, in advance, exactly which excuses he's going to need. When we bring up our measuring instruments, he knows that he'll have to excuse the lack of any carbon dioxide in the air. When we bring in a bag of flour, the claimant knows that he'll need to excuse the lack of any dragon-shaped form in the floury air.

To a Bayesian, a hypothesis isn't something you assert in a loud, emphatic voice. A hypothesis is something that controls your *anticipations*, the probabilities you assign to future experiences. That's what a probability *is*, to a Bayesian—that's what you score, that's what you calibrate. So while our claimant may say loudly, emphatically, and honestly that he *believes* there's an invisible dragon in the garage, he does not *anticipate*

there's an invisible dragon in the garage—he anticipates exactly the same experience as the skeptic.

When I judge the predictions of a hypothesis, I ask which experiences I would anticipate, not which facts I would believe.

* * *

The flip side:

I recently argued with a friend of mine over a question of evolutionary theory. My friend alleged that the clustering of changes in the fossil record (apparently, there are periods of comparative stasis followed by comparatively sharp changes; itself a controversial observation known as "punctuated equilibrium") showed that there was something wrong with our understanding of speciation. My friend thought that there was some unknown force at work, not supernatural, but some natural consideration that standard evolutionary theory didn't take into account. Since my friend didn't give a specific competing hypothesis that produced better predictions, his thesis had to be that the standard evolutionary model was *stupid* with respect to the data—that the standard model made a specific prediction that was wrong; that the model did worse than complete ignorance or some other default competitor.

At first I fell into the trap; I accepted the implicit assumption that the standard model predicted smoothness, and based my argument on my recollection that the fossil record changes weren't as sharp as he claimed. He challenged me to produce an evolutionary intermediate between *Homo erectus* and *Homo sapiens*; I googled and found *Homo heidelbergensis*. He congratulated me and acknowledged that I had scored a major point, but still insisted that the changes were too sharp, and not steady enough. I started to explain why I thought a pattern of uneven change *could* arise from the standard model: environmental selection pressures might not be constant . . . "Aha!" my friend said, "you're making your excuses in advance."

But suppose that the fossil record instead showed a smooth and gradual set of changes. Might my friend have argued that the standard model of evolution as a chaotic and noisy process could not account for such smoothness? If it is a scientific sin to claim post facto that our beloved hypothesis predicts the data, should it not be equally a sin to claim post facto that the competing hypothesis is stupid on the data?

If a hypothesis has a *purely* technical model, there is no trouble; we can compute the prediction of the model formally, without informal variables to provide a handle for post facto meddling. But what of semitechnical theories? Obviously a semitechnical theory must produce some good advance predictions about *something*, or else why bother? But *after* the theory is semi-confirmed, can the detractors claim that the data show a problem with the semitechnical theory, when the "problem" is constructed post facto? At the least the detractors must be very specific about what data a confirmed model predicts stupidly,

and why the confirmed model must make (post facto) that stupid prediction. How sharp a change is "too sharp," quantitatively, for the standard model of evolution to permit? Exactly how much steadiness do you think the standard model of evolution predicts? How do you know? Is it too late to say that, after you've seen the data?

When my friend accused me of making excuses, I paused and asked myself which excuses I anticipated needing to make. I decided that my current grasp of evolutionary theory didn't say anything about whether the rate of evolutionary change should be intermittent and jagged, or smooth and gradual. If I hadn't seen the graph in advance, I could not have predicted it. (Unfortunately, I rendered even that verdict after seeing the data . . .) Maybe there are models in the evolutionary family that would make advance predictions of steadiness or variability, but if so, I don't know about them. More to the point, my friend didn't know either.

It is not always wise, to ask the opponents of a theory what their competitors predict. Get the theory's predictions from the theory's best advocates. Just make sure to write down their predictions in advance. Yes, sometimes a theory's advocates try to make the theory "fit" evidence that plainly doesn't fit. But if you find yourself wondering what a theory predicts, ask first among the theory's advocates, and afterward ask the detractors to cross-examine.

Furthermore: Models may include noise. If we hypothesize that the data are trending slowly and steadily upward, but our measuring instrument has an error of 5%, then it does no good to point to a data point that dips below the previous data point, and shout triumphantly, "See! It went down! Down down down! And don't tell me why your theory fits the dip; you're just making excuses!" Formal, technical models often incorporate explicit error terms. The error term spreads out the likelihood density, decreases the model's precision and reduces the theory's score, but the Bayesian scoring rule still governs. A technical model can allow mistakes, and make mistakes, and still do better than ignorance. In our supermarket example, even the precise hypothesis of 51 still bets only 90% of its probability mass on 51; the precise hypothesis claims only that 51 happens nine times out of ten. Ignoring nine 51s, pointing at one case of 82, and crowing in triumph, does not a refutation make. That's not an excuse, it's an explicit advance prediction of a technical model.

The error term makes the "precise" theory vulnerable to a superprecise alternative that predicted the 82. The standard model would also be vulnerable to a precisely ignorant model that predicted a 60% chance of 51 on the round where we saw 82, spreading out the likelihood more entropically on that particular error. No matter how good the theory, science always has room for a higher-scoring competitor. But if you *don't* present a better alternative, if you try only to show that an accepted theory is *stupid* with respect

to the data, that scientific endeavor may be *more* demanding than just replacing the old theory with a new one.

Astronomers recorded the unexplained perihelion advance of Mercury, unaccounted for under Newtonian physics—or rather, Newtonian physics predicted 5557 seconds of arc per century, where the observed amount was 5600 (Brown 2011, 405-414). But should the scientists of that day have junked Newtonian gravitation based on such small, unexplained counterevidence? What would they have used instead? Eventually, Newton's theory of gravitation *was* set aside, after Einstein's General Relativity precisely explained the orbital discrepancy of Mercury and also made successful advance predictions. But there was no way to know *in advance* that this was how things would turn out.

In the nineteenth century there was a persistent anomaly in the orbit of Uranus. People said, "Maybe Newton's law starts to fail at long distances." Eventually some bright fellows looked at the anomaly and said, "Could this be an unknown outer planet?" Urbain Le Verrier and John Couch Adams independently did some scribbling and figuring, using Newton's standard theory—and predicted Neptune's location to within one degree of arc, dramatically *confirming* Newtonian gravitation (405-414).

Only *after* General Relativity precisely produced the perihelion advance of Mercury, did we *know* Newtonian gravitation would never explain it.

* * *

In the *Intuitive Explanation* we saw how Karl Popper's insight that falsification is stronger than confirmation, translates into a Bayesian truth about likelihood ratios. Popper erred in thinking that falsification was *qualitatively different* from confirmation; both are governed by the same Bayesian rules. But Popper's philosophy reflected an important truth about a quantitative difference between falsification and confirmation.

> Popper was profoundly impressed by the differences between the allegedly "scientific" theories of Freud and Adler and the revolution effected by Einstein's theory of relativity in physics in the first two decades of this century. The main difference between them, as Popper saw it, was that while Einstein's theory was highly "risky," in the sense that it was possible to deduce consequences from it which were, in the light of the then dominant Newtonian physics, highly improbable (e.g. that light is deflected towards solid bodies—confirmed by Eddington's experiments in 1919), and which would, if they turned out to be false, falsify the whole theory, nothing could, even in principle, falsify psychoanalytic theories. These latter, Popper came to feel, have more in common with primitive myths than with genuine science. That is to say, he saw that what is apparently the chief source of strength of psychoanal-

> ysis, and the principal basis on which its claim to scientific status is grounded, viz. its capability to accommodate, and explain, every possible form of human behaviour, is in fact a critical weakness, for it entails that it is not, and could not be, genuinely predictive. Psychoanalytic theories by their nature are insufficiently precise to have negative implications, and so are immunised from experiential falsification. . . .
>
> Popper, then, repudiates induction, and rejects the view that it is the characteristic method of scientific investigation and inference, and substitutes falsifiability in its place. It is easy, he argues, to obtain evidence in favour of virtually any theory, and he consequently holds that such "corroboration," as he terms it, should count scientifically only if it is the positive result of a genuinely 'risky' prediction, which might conceivably have been false. For Popper, a theory is scientific only if it is refutable by a conceivable event. Every genuine test of a scientific theory, then, is logically an attempt to refute or to falsify it. . . .
>
> Every genuine scientific theory then, in Popper's view, is prohibitive, in the sense that it forbids, by implication, particular events or occurrences. (Thornton 2002)

On Popper's philosophy, the strength of a scientific theory is not how much it explains, but how much it *doesn't* explain. The virtue of a scientific theory lies not in the outcomes it *permits*, but in the outcomes it *prohibits*. Freud's theories, which seemed to explain everything, *prohibited* nothing.

Translating this into Bayesian terms, we find that the more outcomes a model *prohibits*, the more probability density the model concentrates in the remaining, permitted outcomes. The more outcomes a theory prohibits, the greater the knowledge-content of the theory. The more daringly a theory exposes itself to falsification, the more definitely it tells you which experiences to anticipate.

A theory that can explain *any* experience corresponds to a hypothesis of complete ignorance—a uniform distribution with probability density spread evenly over every possible outcome.

* * *

One of the most famous lessons of science is the case of the *phlogiston theory of chemistry*.

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

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

Phlogiston escaped from burning substances as visible fire. As the phlogiston escaped, the burning substances lost phlogiston and so became ash, the "true material." Flames extinguished in closed containers because the air became saturated with phlogiston. Charcoal left little residue upon burning because it was nearly pure phlogiston.[4]

This was a more primitive age of science, and so people did not notice and take offense that phlogiston theory made no advance predictions. Instead phlogiston theory just added on more and more independent clauses to explain more and more chemical observations. You couldn't use phlogiston theory to predict the outcome of a chemical transformation—first you looked at the result, then you used phlogiston to explain it. It was not that, having never tried burning a flame in a closed container, phlogiston theorists predicted that the flame would go out when the air became "saturated" with phlogiston. Rather they lit a flame in a container, watched it go out, then said, "The air must have become saturated with phlogiston."

You couldn't even use phlogiston theory to constrain chemical transformations, to say what you did *not* expect to see. Phlogiston theory was infinitely flexible. In excusing everything, it explained nothing; a disguised hypothesis of zero knowledge.

The word *phlogiston* functioned not as an *anticipation-controller* but as a *curiosity-stopper*. You said "Why?" and the answer was "Phlogiston."

* * *

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

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

4. See (Moore 1961); see also Phlogiston Theory, Demise of Phlogiston, and Friedrich Wöhler.

> were coming once more to the acceptance of something and that was a vital principle.
>
> —Lord Kelvin (quoted in Zapato 2008)

This was the theory of *vitalism*; that the difference between living matter and non-living matter consisted of an *élan vital* or *vis vitalis*. Élan vital infused living matter and caused it to move as consciously directed. Élan vital participated in chemical transformations which no mere non-living particles could undergo. Wöhler's artificial synthesis of urea, a component of urine, was a major blow to the vitalistic theory because it showed that "mere chemistry" could duplicate a product of biology (Moore 1961).

Building on the previous lesson of phlogiston, we note at once that *élan vital* functions not as an anticipation-controller but as a curiosity-stopper. Vitalism doesn't explain how the hand moves, nor tell you what transformations to expect from organic chemistry, and vitalism certainly permits no quantitative calculations. "Why? Élan vital!" And that was all there was to vitalism.

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

I quote Lord Kelvin to show that in every generation, there are scientific puzzles so wonderfully *mys-TER-i-ous* that they become sacred, making a *solution* sacrilege. Science is only good for explaining non-mysterious phenomena, like the course of planets, or the transformations of materials, or the biology of life; science can never answer questions about *real* mysteries like consciousness. Surely, if it were possible for science to explain consciousness, it would already have done so? As if all these other matters had not been mysteries for thousands of years and millions of years, from the dawn of intelligent thought right up until science solved them.

People have no sense of history. They learn about stars and chemistry and biology in school and it seems that these matters have always been the proper meat of science, that they have *never been* mysterious. Astrologers and alchemists and vitalists were merely fools, to make such big deals out of such simple questions. When science must deal with some new puzzling phenomenon, it is a great shock to the children of that generation, for they have never encountered something that *feels* mysterious before. Surely such a sacred mystery as consciousness is infinitely beyond the reach of dry scientific thinking; science is only suited to mundane questions such as biology.

Vitalism shared with phlogiston the error of *encapsulating the mystery as a substance*. Fire was mysterious, and the phlogiston theory encapsulated the mystery in a mysterious substance called "phlogiston." Life was a sacred mystery, and vitalism encapsulated

the sacred mystery in a mysterious substance called "élan vital." Neither "explanation" helped concentrate the model's probability density. The "explanation" just wrapped up the question as a small, hard, opaque black ball. In a play written by the author Molière, a physician explains the power of a soporific by claiming that the soporific contains a "dormitive potency"—a fine parody of the art of fake explanation (quoted in Kuhn 1962).

It is a failure of human psychology that, faced with a mysterious phenomenon, we more readily postulate mysterious inherent substances than complex underlying processes.

But the deeper failure is supposing that an *answer* can be mysterious. Mystery is a property of questions, not answers. If a phenomenon feels mysterious, that is a fact about our state of knowledge, not a fact about the phenomenon itself. The vitalists saw a mysterious gap in their knowledge, and postulated a mysterious stuff that plugged the gap. They mixed up the map with the territory. All confusion and dismay exist in the mind, not in reality.

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

* * *

The flip side:

Beware of checklist thinking: Having a *sacred* mystery, or a mysterious answer, is not the same as refusing to explain something. Some elements in our physics are taken as "fundamental," not yet further reduced or explained. But these fundamental elements of our physics are governed by clearly defined, mathematically simple, formally computable causal rules.

Occasionally some crackpot objects to modern physics on the grounds that it does not provide an "underlying mechanism" for a mathematical law currently treated as fundamental. (Claiming that a mathematical law lacks an "underlying mechanism" is one of the entries on the *Crackpot Index* by John Baez [1998].) The "underlying mechanism" the crackpot proposes in answer is vague, verbal, and yields no increase in predictive power—otherwise we would not classify the claimant as a crackpot.

Our current physics makes the electromagnetic field fundamental, and refuses to explain it further. But the "electromagnetic field" is a fundamental governed by clear mathematical rules, with no properties outside the mathematical rules, subject to formal computation to describe its causal effect upon the world. Someday someone may suggest improved math that yields better predictions, but I would not indict the current model on grounds of mysteriousness. A theory that includes *fundamental elements* is not the same as a theory that contains *mysterious elements*.

Fundamentals should be simple. "Life" is not a good fundamental, "oxygen" is a good fundamental, and "electromagnetic field" is a better fundamental. Life might look simple to a vitalist—it's the simple, magical ability of your muscles to move under your mental direction. Why shouldn't life be explained by a simple, magical fundamental substance like *élan vital*? But phenomena that seem *psychologically* very simple—little dots of light in the sky, orangey-bright hot flame, flesh moving under mental direction—often conceal vast depths of underlying complexity. The proposition that life is a complex phenomenon may seem incredible to the vitalist, staring at a blankly opaque mystery with no obvious handles; but yes, Virginia, there is underlying complexity. The criterion of simplicity that is relevant to Occam's Razor is *mathematical* or *computational* simplicity. Once we render down our model into mathematically simple fundamental elements, not in themselves sharing the mysterious qualities of the mystery, interacting in clearly defined ways to produce the formerly mysterious phenomenon as a detailed prediction, that is as non-mysterious as humanity has ever figured out how to make anything.

* * *

The failures of phlogiston and vitalism are historical hindsight. Dare I step out on a limb and name some current theory, not yet disproven, that I think is analogously flawed to vitalism and phlogiston? I shall dare, but don't try this at home. I also warn my readers that they should not accept this opinion of mine with the same confidence that attaches to science's dismissal of phlogiston.

I name the fad of *emergence* or *emergent phenomena*—systems which exhibit high-level behaviors that arise or "emerge" from the interaction of many low-level elements. Taken literally, that description fits every phenomenon in our universe above the level of individual quarks, which is part of the problem.

In decrying the emergence fad, I decry the use of "emergence" *as an explanation in itself*. It's okay to have a completed model to which an emergence enthusiast could attach "emergent" as an adjective. One might legitimately have some *specific* model of how the behavior of an ant colony *emerges* from the behavior of the ants. A hypothesis like that can be formal and/or technical. The model of the ant colony has internal moving parts and produces specific predictions; it's just that the model happens to fit the verbal term "emergent"—the behavior which emerges from modeling many interacting elements is

different from the behavior of those elements considered in isolation. I do not consider it stupid to say that Phenomenon X *emerges from* Y, where Y is some specific model. The phrase "emerges from" is okay, if the phrase precedes some specific model to be judged on its own merits.

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

To say that intelligence is an "emergent phenomenon" fits every possible behavior that intelligence could show, and therefore explains nothing. The model has no moving parts and does not concentrate its probability mass into specific outcomes. It is a disguised hypothesis of zero knowledge.

To see why I object to the academic fad in "emergence," even though I have admitted the legitimacy of the phrase "emerges from," consider that "arises from" is also a legitimate phrase. Gravity arises from the curvature of spacetime (according to a certain specific mathematical model, Einstein's General Relativity). Chemistry arises from interactions between atoms (according to the specific model of quantum electrodynamics). Now suppose I should say that gravity is explained by "arisence" or that chemistry is an "arising phenomenon," and claim that as my explanation.

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

> *Before:* Human intelligence is an emergent product of neurons firing.
> *After:* Human intelligence is a product of neurons firing.
>
> *Before:* The behavior of the ant colony is the emergent outcome of the interactions of many individual ants.
> *After:* The behavior of the ant colony is the outcome of the interactions of many individual ants.
> *Even better:* A colony is made of ants. We can successfully predict some

> aspects of colony behavior using models that include only individual ants, without any global colony variables, showing that we understand how those colony behaviors arise from ant behaviors.

Another good exercise is to replace the word "emergent" with the old word, the explanation that people had to use before emergence was invented.

> *Before:* Life is an emergent phenomenon.
> *After:* Life is a magical phenomenon.

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

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

Magic is unpopular nowadays, unfashionable, not something you could safely postulate in a peer-reviewed journal. Why? Once upon a time, a few exceptionally wise scientists noticed that explanations which invoked "magic" just didn't work as a way of understanding the world. By dint of strenuous evangelism, these wise scientists managed to make magical explanations unfashionable within a small academic community. But humans are still humans, and they have the same emotional needs and intellectual vulnerabilities. So later academics invented a new word, "emergence," that carried exactly the same information content as "magic," but had not yet become unfashionable. "Emergence" became very popular, just as saying "magic" used to be very popular. "Emergence" has the same deep appeal to human psychology, for the same reason. "Emergence" is such a wonderfully easy explanation, and it feels good to say it; it gives you a sacred mystery to worship. Emergence is a popular fad *because* it is the junk food of curiosity. You can explain anything using emergence, and so people do just that; for it feels so wonderful to explain things. Humans are still humans, even if they've taken a few science classes in college. Once they find a way to escape the shackles of settled science, they get up to the same shenanigans as their ancestors, dressed in different clothes but still the same species psychology.

* * *

Many people in this world believe that after dying they will face a stern-eyed fellow named St. Peter, who will examine their actions in life and accumulate a score for morality. Presumably St. Peter's scoring rule is unique and invariant under trivial changes of perspective. Unfortunately, believers cannot obtain a quantitative, precisely computable specification of the scoring rule, which seems rather unfair.

The religion of *Bayesianity* holds that your eternal fate depends on the probability judgments you made in life. Unlike lesser faiths, Bayesianity can give a quantitative, precisely computable specification of how your eternal fate is determined.

Our proper Bayesian scoring rule provides a way to accumulate scores across experiments, and the score is invariant regardless of how we slice up the "experiments" or in what order we accumulate the results. We add up the logarithms of the probabilities. This corresponds to multiplying together the probability assigned to the outcome in each experiment, to find the joint probability of all the experiments together. We take the logarithm to simplify our intuitive understanding of the accumulated score, to maintain our grip on the tiny fractions involved, and to ensure we maximize our *expected* score by stating our honest probabilities rather than placing all our play money on the most probable bet.

Bayesianity states that, when you die, Pierre-Simon Laplace examines every single event in your life, from finding your shoes next to your bed in the morning, to finding your workplace in its accustomed spot. Every losing lottery ticket means you cared enough to play. Laplace assesses the advance probability you assigned to each event. Where you did not assign a precise numerical probability in advance, Laplace examines your degree of anticipation or surprise, extrapolates other possible outcomes and your extrapolated reactions, and renormalizes your extrapolated emotions to a likelihood distribution over possible outcomes. (Hence the phrase "Laplacian superintelligence.")

Then Laplace takes every event in your life, and every probability you assigned to each event, and multiplies all the probabilities together. This is your Final Judgment—the probability you assigned to your life.

Those who follow Bayesianity strive all their lives to maximize their Final Judgment. This is the sole virtue of Bayesianity. The rest is just math.

Mark you: the path of Bayesianity is strict. What probability shall you assign each morning, to the proposition, "The sun shall rise?" (We shall discount such quibbles as cloudy days, and that the Earth orbits the Sun.) Perhaps one who did not follow Bayesianity would be humble, and give a probability of 99.9%. But we who follow Bayesianity shall discard all considerations of modesty and arrogance, and scheme only to maximize our Final Judgment. Like an obsessive video-game player, we care only about this numerical score. We're going to face this Sun-shall-rise issue 365 times per year, so we might be able to improve our Final Judgment considerably by tweaking our probability assignment.

As it stands, even if the Sun rises every morning, every year our Final Judgment will decrease by a factor of 0.7 (0.999^{365}), roughly -0.52 bits. Every two years, our Final Judgment will decrease more than if we found ourselves ignorant of a coinflip's outcome! Intolerable. If we increase our daily probability of sunrise to 99.99%, then each year our

Final Judgment will decrease only by a factor of 0.964. Better. Still, in the unlikely event that we live exactly 70 years and then die, our Final Judgment will only be 7.75% of what it might have been. What if we assign a 99.999% probability to the sunrise? Then after 70 years, our Final Judgment will be multiplied by 77.4%.

Why not assign a probability of 1.0?

One who follows Bayesianity will *never* assign a probability of 1.0 to *anything*. Assigning a probability of 1.0 to some outcome uses up *all* your probability mass. If you assign a probability of 1.0 to some outcome, and reality delivers a different answer, you must have assigned the *actual* outcome a probability of *zero*. This is Bayesianity's sole mortal sin. Zero times anything is zero. When Laplace multiplies together all the probabilities of your life, the combined probability will be zero. Your Final Judgment will be doodly-squat, zilch, nada, nil. No matter how rational your guesses during the rest of your life, you'll spend eternity next to some guy who believed in flying saucers and got all his information from the Weekly World News. Again we find it helpful to take the logarithm, revealing the innocent-sounding "zero" in its true form. Risking an outcome probability of zero is like accepting a bet with a payoff of negative infinity.

What if humanity decides to take apart the Sun for mass (stellar engineering), or to switch off the Sun because it's wasting entropy? Well, you say, you'll see that coming, you'll have a chance to alter your probability assignment before the actual event. What if an Artificial Intelligence in someone's basement recursively self-improves to superintelligence, stealthily develops nanotechnology, and one morning *it* takes apart the Sun? If on the last night of the world you assign a probability of 99.999% to tomorrow's sunrise, your Final Judgment will go down by a factor of 100,000. Minus 50 decibels! Awful, isn't it?

So what is your best strategy? Well, suppose you 50% anticipate that a basement-spawned AI superintelligence will disassemble the Sun sometime in the next ten years, and you figure there's about an equal chance of this happening on any given day between now and then. On any given night, you would 99.98% anticipate the sun rising tomorrow. If this is really what you anticipate, then you have no motive to say anything except 99.98% as your probability. If you feel nervous that this anticipation is too low, or too high, it must not be what you anticipate after your nervousness is taken into account.

But the deeper truth of Bayesianity is this: you cannot game the system. You cannot give a humble answer, nor a confident one. You must figure out exactly how much you anticipate the Sun rising tomorrow, and say that number. You must shave away every hair of modesty or arrogance, and ask whether you expect to end up being scored on the Sun rising, or failing to rise. Look not to your excuses, but ask which excuses you expect to need. After you arrive at your exact degree of anticipation, the only way to further improve your Final Judgment is to improve the accuracy, calibration, and discrimination

of your anticipation. You cannot do better except by guessing better and anticipating more precisely.

Er, well, except that you could commit suicide when you turned five, thereby preventing your Final Judgment from decreasing any further. Or if we patch a new sin onto the utility function, enjoining against suicide, you could flee from mystery, avoiding all situations in which you thought you might not know everything. So much for that religion.

* * *

Ideally, we predict the outcome of the experiment in advance, using our model, and then we perform the experiment to see if the outcome accords with our model. Unfortunately, we can't always control the information stream. Sometimes Nature throws experiences at us, and by the time we think of an explanation, we've already seen the data we're supposed to explain. This was one of the scientific sins committed by nineteenth century evolutionism; Darwin observed the similarity of many species, and their adaptation to particular local environments, before the hypothesis of natural selection occurred to him. Nineteenth century evolutionism began life as a post facto explanation, not an advance prediction.

Nor is this a trouble only of semitechnical theories. In 1846, the successful deduction of Neptune's existence from gravitational perturbations in the orbit of Uranus was considered a grand triumph for Newton's theory of gravitation. Why? Because Neptune's existence was the first observation that confirmed an *advance* prediction of Newtonian gravitation. All the other phenomena that Newton explained, such as orbits and orbital perturbations and tides, had been observed in great detail before Newton explained them. No one seriously doubted that Newton's theory was correct. Newton's theory explained too much too precisely, and it replaced a collection of ad hoc models with a single unified mathematical law. Even so, the advance prediction of Neptune's existence, followed by the observation of Neptune at almost exactly the predicted location, was considered the first grand triumph of Newton's theory at predicting what no previous model could predict. Considerable time elapsed between widespread acceptance of Newton's theory and the first impressive *advance* prediction of Newtonian gravitation. By the time Newton came up with his theory, scientists had already observed, in great detail, most of the phenomena that Newtonian gravitation predicted.

But the rule of advance prediction is a morality of science, not a law of probability theory. If you have already seen the data you must explain, then Science may darn you to heck, but your predicament doesn't collapse the laws of probability theory. What does happen is that it becomes much more difficult for a hapless human to *obey* the laws of probability theory. When you're deciding how to rate a hypothesis according to the Bayesian scoring rule, you need to figure out how much probability mass that

hypothesis assigns to the observed outcome. If we must make our predictions in advance, then it's easier to notice when someone is trying to claim every possible outcome as an advance prediction, using too much probability mass, being deliberately vague to avoid falsification, and so on.

No numerologist can predict next week's winning lottery numbers, but they will be happy to explain the mystical significance of last week's winning lottery numbers. Say the winning Mega Ball was seven in last week's lottery, out of 52 possible outcomes. Obviously this happened because seven is the lucky number. So will the Mega Ball in next week's lottery also come up seven? We understand that it's not certain, of course, but if it's the lucky number, you ought to assign a probability of higher than 1/52 . . . and then we'll score your guesses over the course of a few years, and if your score is too low we'll have you flogged . . . what's that you say? You want to assign a probability of exactly 1/52? But that's the same probability as every other number; what happened to seven being lucky? No, sorry, you can't assign a 90% probability to seven and also a 90% probability to eleven. We understand they're both lucky numbers. Yes, we understand that they're *very* lucky numbers. But that's not how it works.

Even if the listener does not know the way of Bayes and does not ask for formal probabilities, they will probably become suspicious if you try to cover too many bases. Suppose they ask you to predict next week's winning Mega Ball, and you use numerology to explain why the number one ball would fit your theory very well, and why the number two ball would fit your theory very well, and why the number three ball would fit your theory very well . . . even the most credulous listener might begin to ask questions by the time you got to twelve. Maybe you could tell us which numbers are unlucky and definitely won't win the lottery? Well, thirteen is unlucky, but it's not absolutely *impossible* (you hedge, *anticipating* in advance which excuse you might need).

But if we ask you to explain *last week's* lottery numbers, why, the seven was practically inevitable. That seven should definitely count as a major success for the "lucky numbers" model of the lottery. And it couldn't possibly have been thirteen; luck theory rules that straight out.

* * *

Imagine that you wake up one morning and your left arm has been replaced by a blue tentacle. The blue tentacle obeys your motor commands—you can use it to pick up glasses, drive a car, etc. How would you explain this hypothetical scenario? Take a moment to ponder this puzzle before continuing.

spoiler space
spoiler space
spoiler space

How would I explain the event of my left arm being replaced by a blue tentacle? The answer is that I wouldn't. It isn't going to happen.

It would be easy enough to produce a verbal explanation that "fit" the hypothetical. There are many explanations that can "fit" anything, including (as a special case of "anything") my arm being replaced by a blue tentacle. Divine intervention is a good all-purpose explanation. Or aliens with arbitrary motives and capabilities. Or I could be mad, hallucinating, dreaming my life away in a hospital. Such explanations "fit" all outcomes equally well, and equally poorly, equating to hypotheses of complete ignorance.

The test of whether a model of reality "explains" my arm turning into a blue tentacle, is whether the model concentrates significant probability mass into that *particular* outcome. Why that dream, in the hospital? Why would aliens do that particular thing to me, as opposed to the other billion things they might do? Why would my arm turn into a tentacle on that morning, after remaining an arm through every other morning of my life? And in all cases I must look for an argument compelling enough to make that particular prediction in *advance*, not mere compatibility. Once I already knew the outcome, it would become far more difficult to sift through hypotheses to find good explanations. Whatever hypothesis I tried, I would be hard-pressed not to allocate more probability mass to yesterday's blue-tentacle outcome than if I extrapolated blindly, seeking the model's *most* likely prediction for tomorrow.

A model does not always predict all the features of the data. Nature has no privileged tendency to present me with solvable challenges. Perhaps a deity toys with me, and the deity's mind is computationally intractable. If I flip a fair coin there is no way to further explain the outcome, no model that makes a better prediction than the maximum-entropy hypothesis. But if I guess a model with no internal detail or a model that makes no further predictions, I not only have no reason to believe that guess, I have no reason to care. Last night my arm was replaced with a blue tentacle. Why? Aliens! So what will they do tomorrow? Similarly, if I attribute the blue tentacle to a hallucination as I dream my life away in a coma, I still don't know any more about what I'll hallucinate tomorrow. So why do I care whether it was aliens or hallucination?

What might be a *good* explanation, then, if I woke up one morning and found my arm transformed into a blue tentacle? To claim a "good explanation" for this hypothetical experience would require an argument such that, contemplating the hypothetical argument *now*, *before* my arm has transformed into a blue tentacle, I would go to sleep worrying that my arm *really would* transform into a tentacle.

People play games with plausibility, explaining events they expect to never actually encounter, yet this necessarily violates the laws of probability theory. How many people who thought they could "explain" the hypothetical experience of waking up with their arm replaced by a tentacle, would go to sleep wondering if it might really happen to

them? Had they the courage of their convictions, they would say: I do not expect to ever encounter this hypothetical experience, and therefore I cannot explain, nor have I a motive to try. Such things only happen in webcomics, and I need not prepare explanations, for in real life I shall never have a chance to use them. If I ever find myself in this impossible situation, let me miss no jot or tittle of my valuable bewilderment.

To a Bayesian, probabilities are anticipations, not mere beliefs to proclaim from the rooftops. If I have a model that assigns probability mass to waking up with a blue tentacle, then I am nervous about waking up with a blue tentacle. What if the model is a fanciful one, like a witch casting a spell that transports me into a randomly selected webcomic? Then the *prior probability* of webcomic witchery is so low that my *real-world* understanding doesn't assign any significant weight to that hypothesis. The witchcraft hypothesis, if taken as a given, might assign non-insignificant likelihood to waking up with a blue tentacle. But my anticipation of that hypothesis is so low that I don't anticipate any of the predictions of that hypothesis. That I can conceive of a witchcraft hypothesis should in no wise diminish my stark bewilderment if I actually wake up with a tentacle, because the real-world probability I assign to the witchcraft hypothesis is effectively zero. My zero-probability hypothesis wouldn't help me *explain* waking up with a tentacle, because the argument isn't good enough to make me *anticipate* waking up with a tentacle.

In the laws of probability theory, likelihood distributions are fixed properties of a hypothesis. In the art of rationality, to *explain* is to *anticipate*. To *anticipate* is to *explain*. Suppose I am a medical researcher, and in the ordinary course of pursuing my research, I notice that my clever new theory of anatomy seems to permit a small and vague possibility that my arm will transform into a blue tentacle. "Ha ha!" I say, "how remarkable and silly!" and feel ever so slightly nervous. *That* would be a good explanation for waking up with a tentacle, if it ever happened.

If a chain of reasoning doesn't make me nervous, in advance, about waking up with a tentacle, then that reasoning would be a poor explanation if the event *did* happen, because the combination of prior probability and likelihood was too low to make me allocate any significant real-world probability mass to that outcome.

If you start from well-calibrated priors, and you apply Bayesian reasoning, you'll end up with well-calibrated conclusions. Imagine that two million entities, scattered across different planets in the universe, have the opportunity to encounter something so strange as waking up with a tentacle (or—gasp!—ten fingers). One million of these entities say "one in a thousand" for the prior probability of some hypothesis X, and each hypothesis X says "one in a hundred" for the likelihood of waking up with a tentacle. And one million of these entities say "one in a hundred" for the prior probability of some hypothesis Y, and each hypothesis Y says "one in ten" for the likelihood of waking up with a tentacle.

If we suppose that all entities are well-calibrated, then we shall look across the universe and find ten entities who wound up with a tentacle because of hypotheses of plausibility class *X*, and a thousand entities who wound up with tentacles because of hypotheses of plausibility class *Y*. So if you find yourself with a tentacle, and *if* your probabilities are well-calibrated, then the tentacle is more likely to stem from a hypothesis you would class as probable than a hypothesis you would class as improbable. (What if your probabilities are poorly calibrated, so that when you say "million-to-one" it happens one time out of twenty? Then you're grossly overconfident, and we adjust your probabilities in the direction of less discrimination and greater entropy.)

The hypothesis of being transported into a webcomic, even if it "explains" the scenario of waking up with a blue tentacle, is a poor explanation because of its low prior probability. The webcomic hypothesis doesn't contribute to explaining the tentacle, because it doesn't make you anticipate waking up with a tentacle.

If we start with a quadrillion sentient minds scattered across the universe, quite a lot of entities will encounter events that are very likely, only about a mere million entities will experience events with lifetime likelihoods of a billion-to-one (as we would anticipate, surveying with infinite eyes and perfect calibration), and not a single entity will experience the impossible.

If, somehow, you really did wake up with a tentacle, it would likely be because of something much more probable than "being transported into a webcomic," some perfectly normal reason to wake up with a tentacle which you just didn't see coming. A reason like what? I don't know. Nothing. I don't anticipate waking up with a tentacle, so I can't give any good explanation for it. Why should I bother crafting excuses that I don't expect to use? If I was worried I might someday need a clever excuse for waking up with a tentacle, the *reason I was nervous about the possibility* would be *my* explanation.

Reality dishes out experiences using probability, not plausibility. If you find out that your laptop doesn't obey conservation of momentum, then reality must think that a perfectly normal thing to do to you. How could violating conservation of momentum possibly be perfectly normal? I anticipate that question has no answer and will never need answering. Similarly, people do *not* wake up with tentacles, so apparently it is *not* perfectly normal.

* * *

There is a shattering truth, so surprising and terrifying that people resist the implications with all their strength. Yet there are a lonely few with the courage to accept this satori. Here is wisdom, if you would be wise:

Since the beginning
Not one unusual thing
Has ever happened.

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.

References

Baez, John. 1998. "The Crackpot Index." Accessed August 25, 2012. http://math.ucr.edu/home/baez/crackpot.html.

Brown, Kevin. 2011. *Reflections On Relativity*. Raleigh, NC: printed by author. http://www.mathpages.com/rr/rrtoc.htm.

Darwin, Charles. 1859. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. 1st ed. London: John Murray. http://darwin-online.org.uk/content/frameset?viewtype=text&itemID=F373&pageseq=1.

———. 1874. *The Descent of Man, and Selection in Relation to Sex*. 2nd ed. London: John Murray. http://darwin-online.org.uk/content/frameset?itemID=F944&viewtype=text&pageseq=1.

Dawes, Robyn M. 1988. *Rational Choice in An Uncertain World*. 1st ed. Edited by Jerome Kagan. San Diego, CA: Harcourt Brace Jovanovich.

Feynman, Richard P., Robert B. Leighton, and Matthew L. Sands. 1963. *The Feynman Lectures on Physics*. 3 vols. Reading, MA: Addison-Wesley.

Imagination Engines, Inc. 2011. "The Imagination Engine® or Imagitron™." http://www.imagination-engines.com/ie.htm.

Jaynes, E. T. 2003. *Probability Theory: The Logic of Science*. Edited by G. Larry Bretthorst. New York: Cambridge University Press. doi:10.2277/0521592712.

Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. 1st ed. Chicago: University of Chicago Press.

Moore, Ruth E. 1961. *The Coil Of Life: The Story Of The Great Discoveries In The Life Sciences*. 1st ed. New York: Alfred A. Knopf.

Pirsig, Robert M. 1974. *Zen and the Art of Motorcycle Maintenance: An Inquiry Into Values*. 1st ed. New York: Morrow.

Popper, Karl R. 1959. *The Logic of Scientific Discovery*. New York: Basic Books.

Robinson, Dave, and Judy Groves. 1998. *Philosophy for Beginners*. 1st ed. Cambridge: Icon Books.

Sagan, Carl. 1995. *The Demon-Haunted World: Science as a Candle in the Dark*. 1st ed. New York: Random House.

Schul, Yaacov, and Ruth Mayo. 2003. "Searching for Certainty in an Uncertain World: The Difficulty of Giving Up the Experiential for the Rational Mode of Thinking." *Journal of Behavioral Decision Making* 16 (2): 93–106. doi:10.1002/bdm.434.

Spee, Friedrich. (1631) 2003. *Cautio Criminalis, or a Book on Witch Trials*. Edited and translated by Marcus Hellyer. Studies in Early Modern German History. Charlottesville: University of Virginia Press.

TalkOrigins Foundation. 2012. "Frequently Asked Questions About Creationism and Evolution." Accessed August 25. http://www.talkorigins.org/origins/faqs-qa.html.

Thornton, Stephen. 2002. "Karl Popper." In *The Stanford Encyclopedia of Philosophy*, Winter 2002, edited by Edward N. Zalta. Stanford University. http://plato.stanford.edu/archives/win2002/entries/popper/.

Tversky, Amos, and Ward Edwards. 1966. "Information Versus Reward in Binary Choices." *Journal of Experimental Psychology* 71 (5): 680–683. doi:10.1037/h0023123.

Verhagen, Joachim, ed. 2001. In *Science Jokes,* Version 7.27. October 27. http://web.archive.org/web/20060424082937/http://www.nvon.nl/scheik/best/diversen/scijokes/scijokes.txt.

Williams, George C. 1966. *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought.* Princeton Science Library. Princeton, NJ: Princeton University Press.

Yates, J. Frank, Ju-Whei Lee, Winston R. Sieck, Incheol Choi, and Paul C. Price. 2002. "Probability Judgment Across Cultures." In *Heuristics and Biases: The Psychology of Intuitive Judgment,* edited by Thomas Gilovich, Dale Griffin, and Daniel Kahneman, 271–291. New York: Cambridge University Press. doi:10.2277/0521796792.

Yudkowsky, Eliezer. 2003. "An Intuitive Explanation of Bayes' Theorem." Unpublished manuscript. Last revised June 4, 2006. http://yudkowsky.net/rational/bayes.

Zapato, Lyle. 2008. "Lord Kelvin Quotations." December 14. Accessed August 25, 2012. http://zapatopi.net/kelvin/quotes/.