

# Information Theory for Intelligent People

Simon DeDeo\*

September 9, 2018

## Contents

|           |  |           |
|-----------|--|-----------|
| <b>1</b>  | <b>Twenty Questions</b>  | <b>1</b>  |
| <b>2</b>  | <b>Sidebar: Information on Ice</b>   | <b>4</b>  |
| <b>3</b>  | <b>Encoding and Memory</b>   | <b>4</b>  |
| <b>4</b>  | <b>Coarse-graining</b>   | <b>5</b>  |
| <b>5</b>  | <b>Alternatives to Entropy?</b>  | <b>7</b>  |
| <b>6</b>  | <b>Coding Failure, Cognitive Surprise, and Kullback-Leibler Divergence</b> | <b>7</b>  |
| <b>7</b>  | <b>Einstein and Cromwell’s Rule</b>  | <b>10</b> |
| <b>8</b>  | <b>Mutual Information</b>  | <b>10</b> |
| <b>9</b>  | <b>Jensen-Shannon Distance</b>   | <b>11</b> |
| <b>10</b> | <b>A Note on Measuring Information</b>                                     | <b>12</b> |
| <b>11</b> | <b>Minds and Information</b>   | <b>13</b> |

## 1 Twenty Questions

The story of information theory begins with the children’s game usually known as “twenty questions”. The first player (the “adult”) in this two-player game thinks of something, and by a series of yes-no questions, the other player (the “child”) attempts to guess what it is. “Is it bigger than a breadbox?” No. “Does it have fur?” Yes. “Is it a mammal?” No. And so forth.

If you play this game for a while, you learn that some questions work better than others. Children usually learn that it’s a good idea to eliminate general categories first before becoming

---

\*Article modified from text for the Santa Fe Institute Complex Systems Summer School 2012, and updated for a meeting of the Indiana University Center for 18th Century Studies in 2015, a Colloquium at Tufts University Department of Cognitive Science on Play and String Theory in 2016, and a meeting of the SFI ACTioN Network in 2017. Please send corrections, comments and feedback to [sdedeo@andrew.cmu.edu](mailto:sdedeo@andrew.cmu.edu); <http://santafe.edu/~simon>.

more specific, for example. If you ask on the first round “is it a carburetor?” you are likely wasting time—unless you’re playing the game on Car Talk.

If a game is lasting a *very* long time, you might start to wonder if you could improve your strategy beyond this simple rule of thumb. “Could I have gotten the answer sooner, if I had skipped that useless question about the fur?” A moment’s reflection shows that, in fact, what counts as a good set of questions depends upon the player: if someone is biased towards material things, you’ll tend to focus on questions that split hairs among weights, sizes, and shapes. If someone is biased towards historical figures, you might split hairs about eras of birth.

Strange as it may seem at first, it turns out that for any particular opponent, you can talk not just about strategies being better or worse, but about the possibility of an optimal strategy, one that is better (or at least as good as) any other you might invent. To see how this works, first imagine writing down a script for playing against an opponent, a set of rules that covers all the contingencies: “first ask if it’s a person; then if yes, ask if they were born before 1900, if no, ask if it’s a country...”; or for another friend: “first ask if it’s bigger than a breadbox; if yes, then...”; or for someone else, “first ask if it’s Donald Duck. It almost always is.” Visually, the script could be represented by a branching tree (Fig. 1), with a question at each branch, and one of two paths to take depending on the opponent’s answer.

Once we specify a script, we can try to work out how effective it is. if we describe your opponent by the list of probabilities he has of choosing any particular option, we can then talk about the average number of yes/no questions it takes before the game is finished. Let’s say there are  $N$  words, and label the words  $x_i$ , where the index  $i$  runs from one to  $N$ . For each thing your opponent could be thinking of,  $x_i$ , the game will end in a predetermined number of steps,  $L(x_i)$ —you can just read this off the tree, if you like. The average number of steps is just the sum of all the  $L(x_i)$ , weighted by the probability  $P(x_i)$ ,

$$\text{Script Performance} = \sum_{i=1}^N P(x_i) L(x_i) \quad (1)$$

Having defined a script, and figured out how to compute how quickly it can finish the game, we can then ask the big question: what’s the optimal script, the script with the shortest average time-to-finish for that particular player. Consider a simple case where there are three options—“tree”, “car”, and “bird”; Fig. 1, which you looked at before. If the adult picks “car” half the time when he plays the game, and the other two options a quarter of the time each, then the optimal script for the child is easy to guess. The child should ask, first, if it is a car, because then, half the time, the game will be over in one question. The other half of the time, she will then go on to ask “is it a tree”, and now (knowing that there are only three options) the answer, even if it is “no”, definitively picks out one of the two options. The average number of yes/no questions is 1.5, and a little thought (or experimentation) shows that that’s the best you can do.

To recap: we just figured out the optimal tree, and then, using the probabilities  $P(x)$ , worked out exactly how well it did (1.5 questions, on average). Now things get very strange. It so turns out that we could have gotten the answer to the second question without having to solve the first. In particular, if we compute the quantity  $H(X)$ ,

$$H(X) = - \sum_{i=1}^N P(x_i) \log_2 P(x_i), \quad (2)$$

it is either equal to, or slightly less than, the average length for the optimal script. When it is slightly less than, the true length is always less than or equal to  $H(X)$  rounded up to the nearest integer.

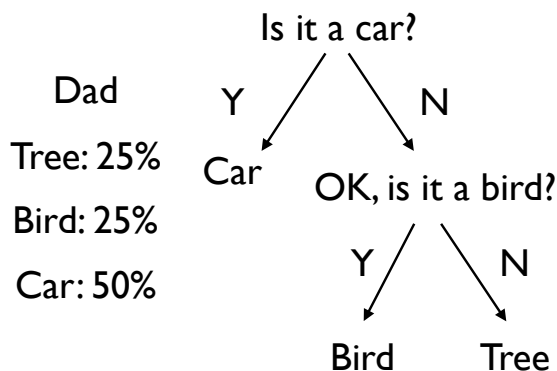


Figure 1: A script for twenty questions, where there are only three options: at each branch-point in the tree, we ask a question, depending on the answer, we take either the left-hand or right-hand branch. Depending on the nature of the opponent, we can then compute the average number of questions required to terminate the game. Given an opponent (“Dad”) who picks “car” half the time, and the two other options one-quarter of the time each, the script here will terminate in one step half the time (“Is it a car?” “Yes”); and in two steps the other half of the time (“Is it a car?” “No.” “OK, is it a bird?” “Yes.”). Since we know that Dad only picks from a set of three, two “no”s suffice to tell us it’s a tree. As we’ll see later, this particular script turns out to be optimal; there’s no set of questions that will end the game sooner (on average).

Here  $\log_2 x$  is the logarithm of  $x$ , in base two; so, for example,  $\log_2 1/8$  is equal to  $-3$  because  $1/8$  is equal to  $2^{-3}$ . You can check by explicit computation that  $H(X)$  for the tree/car/bird problem gets the answer exactly; this is true when the probabilities are all integer powers of  $1/2$ .<sup>1</sup>

Eq. 2 is quite a remarkable result: we’re able to compute the average length for a script without explicitly constructing it. Indeed, there are often a number of different, equivalent optimal scripts (in Fig. 1, for example, the second question could be “is it a tree”), and in many cases, particularly when there are a large number of options, script construction is non-trivial. We won’t prove it in this brief introduction, but if you compare  $H(X)$  to the way you’d compute the average length of an arbitrary tree, Eq. 1, you’ll notice that it suggests that the number of questions to get to choice  $x$ ,  $L(x)$ , is equal to  $-\log_2 P(x)$ —*i.e.*, the less likely possibilities are buried deeper in the question tree.

$H(X)$  is the basic quantity in information theory. In an important sense, all the other quantities we’ll compute are variant on it.  $H(X)$  goes by a number of different names: “uncertainty”, “information”, even “entropy” (a term from the physical sciences, which we’ll return to later). It has units—bits—a name which comes from the phrase “binary digit”, which we’ll understand a bit better below. For any probability distribution, we can now talk about how uncertain we are about the outcome, how much information is in the process, or “how much entropy the process has”, and even (if we are told, or find out, what the probabilities actually are) measure it, in bits.

<sup>1</sup>If this little inaccuracy in the relationship between  $H(X)$  and the optimal script bothers you, here’s a way to solve it. Imagine that instead of building a script for a single round of the game, you build it for  $N$  rounds simultaneously. You’re now asking questions to resolve what the last  $N$  guesses were, all at once, and the questions might be weird—“was your fourth guess an animal and your twelfth guess larger than a breadbox?”—but no matter. The question tree will take longer, but the average number of questions for the simultaneous game, divided by  $N$ , will get closer and closer to  $H(X)$  as  $N$  gets large.

## 2 Sidebar: Information on Ice

$H(X)$  is a fundamentally epistemic quantity. It quantifies how an actual agent in the world goes about gathering information about what is going on. The adult knows with certainty what he has in mind; it is the child, asking yes/no questions, to whom we attach the uncertainty. (To learn more about the modern interpretation of probabilities as describing mental states, rather than facts about the world, see the companion article, Bayesian Reasoning for Intelligent People.)

Because of this, we're tempted to say that whatever information theory measures is a subjective thing, a fact not about the thing, but rather the mind of the beholder. Since we're usually quantifying information in terms of what goes on (literally) in someone's head, this interpretation works very well. But if you want to go a little deeper, you'll discover that information in this subjective sense corresponds directly to a basic quantity in the physical sciences, entropy. Entropy is a fact about the world; you can look up, for example, the "entropy released when ice melts into water". How is this possible?

Let's take the water-ice case. Ice is a crystal, which means that, in any little patch of ice, the water molecules are arranged in a repeating pattern. The nature of that crystal is such that for every position that a water molecule occupies, there's a gap right next door. When ice melts, there are (roughly) the same number of water molecules in that patch. But now they're jumbled about, no longer constrained to stay on the crystal positions, and (in particular) they can now float freely between where they "should" be, and that gap next door. Put in an epistemic fashion, once the ice melts, you become more uncertain about each of the water molecules. You have to ask not only, say, "what's the orientation of the molecule?", but also "is it on the crystal lattice, or off?" That's, literally, one extra yes/no question, one extra bit, and in fact if you do the conversion to the units the chemists use, you get pretty close to the measured value if you say that "the change in entropy when ice melts to water is one bit/molecule"

This is exceptional. A basic quantity in industrial chemistry is directly related to a statement about how physical changes affect what we (or, rather, an ideal angelic agent who can operate on the molecular level) can know. Chemists don't usually talk this way. I learned this way of computing the entropy of the ice-water transition from a 1970 paper published in the Soviet Union, by Yu. V. Gurikov of the All-Union Scientific-Research Institute of Petrochemical Processes, Leningrad. But it is true nonetheless.

So something strange is going on: information is "real." It's up to you to decide, perhaps on the sofa one evening, whether that means that mind is just matter, or (conversely) that matter is an illusion of mind. To give a sense of current exchange rates between mind and matter: the number of bits released when one gram of ice melts is about 100 billion Terabytes, just a little bit more than all the long-term storage humans attach to their computers. I can't tell if it's more amazing that (1) the information in a cube of ice exceeds all of our hard-drives, or (2) humans are starting to get within striking distance of basic thermodynamic limits. Think of that the next time you have a Scotch on the rocks or a well-shaken Cosmopolitan (or lemonade on ice, if you tee-total).

## 3 Encoding and Memory

Let's return to the question script, understood as a binary yes/no tree. Since the answer is arrived at by a specific string of yes/no answers in the script, that string can be understood as an encoding of the answers themselves. In the script of Fig. 1, for example, we can encode "car" as Y, "tree" as NY, and "bird" as NN, a series of "binary digits" (hence, bits) that could be written 0/1 so they look more computery. Each string is unique, and so it really is a code. If two strings match

exactly, it means that the script was unable to distinguish them, and a new question needs to be added. You can then transmit to a friend the guess by flashing the Y/N code—long-flash for Y, say, short-flash for N.

The code has some very nice properties. In particular, you don’t need a special “space” or “pause” character to transmit multiple symbols. If you encode a series of choices in an unbroken fashion, they can be uniquely decoded. Consider the transmission YNYNNYNYNNY. Because the code is variable length, you can’t figure out how many choices are in this list from the length alone. Because a valid codeword is never the prefix to another valid codeword, however, as you read along you know when you’ve finished reading a word and are beginning a new one (try it). Reading left to right gives an unambiguous decomposition. (It’s a nice little puzzle to try to prove this to yourself. Can you see why?)

When the tree is optimal, the encoding is efficient. Computer scientists recognize the construction of the optimal tree as *Huffman Encoding*. Huffman Encoding is the most basic form of lossless compression, when your information units come in chunks—the choices of the adult—and you want to transmit them efficiently in a binary code down, say, a telephone wire.  $H(X)$  then gives you the number of “bits per symbol” your system requires on average; if the underlying process you wish to encode has  $H(X)$  equal to, say, 55, then if the symbols are thrown up at a rate of one per second, you’ll need transmission system that can handle (on average) 55 bits per second. Sometimes you’ll do much better (if you get a common symbol, it will tend to have a short code, and your telegraph operator can take a break) and sometimes much worse (in rare cases, you’re standing around asking questions much longer than you expect).

Huffman Encoding is an incredibly simple way to achieve optimal transmission and storage, and it’s therefore widespread in the computer world. When you “zip” something, the algorithm under the hood is very often some variant of Huffman Encoding. It’s a lossless encoding, meaning that even though it generally finds a much faster way to transmit the information, nothing is lost. It just takes advantage of the biases in the source, giving nicknames or shortcuts to the most common things you say.<sup>2</sup>

It’s more than just an engineering trick, however. The information encoding story enlarges the epistemic “script” interpretation by presenting a story about optimal storage, or communication. The outcome of a process with a lot of uncertainty is harder to remember, or transmit; conversely, an agent with a little experience can develop a codebook that allows them to efficiently gather and store information about the world. We’ll return to these ideas further down when we talk about expectations and learning.

## 4 Coarse-graining

Our question-tree presentation is a nice way to introduce the story of information theory, but it’s not the only way, and, historically, it wasn’t the way it was introduced by the founder of information theory, Claude Shannon, when he was working at AT&T (see Ref. [1]).

Shannon built up his story axiomatically, by saying that he wanted to measure the uncertainty in a process. We want a function, in other words, call it  $H(\vec{p})$ , that takes a list of probabilities and spits out a single number, uncertainty. Let’s require the function to obey four simple axioms. The first two are so simple that they almost seem trivial. The third is intuitive. The fourth is deep.

---

<sup>2</sup>Note that unless you have a pre-arrangement with the other end, you’ll need to transmit the tree itself—the question script—down the wire as well.

1. Continuity (if I only change the probabilities a little, the information of the process should change only a little).
2. Symmetry (if I reorder the list of probabilities I gave you, you should get the same answer).
3. Condition of Maximum Information:  $H(\vec{p})$  is at its maximum value when all the  $p_i$  are equal.
4. Coarse-Graining (discussed below).

Continuity is simple, but Symmetry is a bit loaded: it says that the information is the same whether the probability of events I feed it is (for example)  $\{p_A = 0.2, p_B = 0.8\}$  or  $\{p_A = 0.8, p_B = 0.2\}$ . Of course, if the two probabilities represent the dispositions of a jury, option A is “guilty” and option B is “not guilty,” the prisoner will care a great deal! But  $H$  does not care. For this reason, information is often called a “syntactic” theory, concerned only with the properties of symbols in abstraction from their meanings.

The Maximum condition fits with our folk concept: if every possible outcome (from the list of possible outcomes) is equally likely, the process has the maximum information.<sup>3</sup>

The Coarse-Graining axiom takes the idea of uncertainty step further. It’s really about how we group things together and throw out distinctions. When we say “he was driving a car” rather than “he was driving a red Chevrolet with Massachusetts plates”, we’re coarse-graining, ignoring, or refusing to transmit, a bunch of information that would distinguish very different events. The information you discard may or may not be useful, depending on the use you expect to put the information to, but at the very least it’s more efficient to drop it. Imagine, for example, an indulgent father who allows his child to get “close enough” in the game of twenty questions; depending on the child’s script, it could have a significant effect on the game.

For a more formal definition of coarse graining, let’s say we have a set,  $X$ , with three options,  $\{a, b, c\}$ . I can talk about the uncertainty of a process that spits out one of these three symbols. But what if I don’t care about (or can’t tell) the difference between  $b$  and  $c$ ? Instead of making a finer distinction between  $b$  and  $c$ , I just lump them together into some super-symbol  $S$ . If I coarse-grain  $X$  in this way, I can talk about the uncertainty of probability distribution over the reduced set,  $X'$ ,  $\{a, S\}$ , where  $S$  refers to “ $b$  or  $c$ , I don’t care”, and  $p_{bc}$  is just  $p_b + p_c$ .

When Shannon thought about this process, he realized it could be a very nice way to constrain the function form of entropy. In general, we’d like  $H(X')$  to be less than (or at worst, equal to)  $H(X)$ —if you have an indulgent father, it will help you win faster. Even better, Shannon decided he wanted the following “tree like” property to hold:

$$H(X) = H(X') + p_{bc}H(G), \tag{3}$$

where  $H(G)$  is the uncertainty of the choice between the group  $G$  containing  $b$  and  $c$ ; more explicitly, it’s the uncertainty of the distribution  $\{p_b/p_{bc}, p_c/p_{bc}\}$ .

People reading Shannon’s original paper realized that Eq. 3 was a seriously elegant move, and to see why it’s worth looking at it in different ways. If, for example, you consider a question script, then  $H(X')$  is the average length of the coarse-grained script, and  $H(G)$  the average length of the

---

<sup>3</sup>Note! Do not get confused here. The process itself in this case is very random, and we often associate randomness with lack of information (*e.g.*, if a student starts talking randomly in class, we say he is telling us nothing). But information is concerned with specifying the outcome of a process; imagine having to describe the behavior of a student to a doctor. If the student’s behavior is very random, you have to have a longer conversation (“on Monday he was taking about cats, but then Tuesday he was on to bus schedules, and Wednesday it was about the filling in his teeth...”) as opposed to a less random, and thus lower information, process (“every day, he says hello and that’s it.”)

subscript necessary to make the fine-grained distinctions. The average length of the full script is equal to the coarse-grained script except in the case you have to split within the  $G$ , which happens with probability  $p_{bc}$ . So the coarse-graining axiom contains within itself that branching, tree-like logic.

Once we demand this mathematical property hold, something magical happens. There is now only (up to a constant factor) one (one!) possible mathematical form. It is

$$H(\{p_1, p_2, \dots, p_n\}) = - \sum_{i=1}^n p_i \log p_i \quad (4)$$

where the choice of the base of the logarithm is the one freedom. If you chose the base-2 logarithm,  $H$  has units of bits—the same bits as in the opening question-script story. Any other choice of function, beyond picking different bases, will violate at least one of the four conditions (and usually more than one). In fact, as is easy to show, with this choice, the coarse-graining property of Eq. 3 holds for an arbitrary coarse-graining operation where a fine-grained description  $X$  is transformed into a coarse-grained description  $X'$  with a bunch of different groups  $G_1$ ,  $G_2$ , and so on.

## 5 Alternatives to Entropy?

$H(X)$  can also be considered a measure of diversity; indeed, it's a much better version than simply “counting up appearances of unique types”. In many fields where people have tried to come up with a diversity measure, they've ended up inventing a function that, while not quite equal to  $H(X)$ , approximates it over some interval while not having any of its nice properties. One common alternative choice is to work from the idea that while probabilities must sum to unity, if you have a reasonably even distribution, with lots of small probabilities, the sum of the squares will be lower. Thus, one can define a diversity index

$$D(\{p_1, p_2, \dots, p_n\}) = 1 - \sum_{i=1}^n p_i^2, \quad (5)$$

which is equal to zero when only one option is possible, and gets larger as other options come in to play; another justification for the choice is that it's equal to the probability that two samples taken in a row have different types. It can be shown that  $D(X)$  is proportional to a second-order Taylor series approximation to  $H(X)$ ; see Fig. 2 for the binary cases where there are two probabilities.  $D(X)$  satisfies three of the conditions, but fails to obey the coarse-graining principle.

While  $D(X)$  is nice because you don't have to take logarithms, and so can sometimes be used in algebraic solutions, I recommend not reinventing the wheel!  $H(X)$  has so many beautiful and useful properties—we'll see more later—it's a pity to spoil it. Economists are particularly bad at doing this, and the Herfindahl index (Economics), as well as the Simpson index (Ecology) are both equal to  $1 - D(X)$ ; there's also the Inverse Simpson index,  $1/D(X)$ , which has the same property as  $1 - D(X)$  of being large when probabilities are more “evenly spread”.

## 6 Coding Failure, Cognitive Surprise, and Kullback-Leibler Divergence

An optimal script is, in general, only optimal for a particular probability distribution over symbols. What happens when the distribution changes, but you keep the same script? Let's say that the

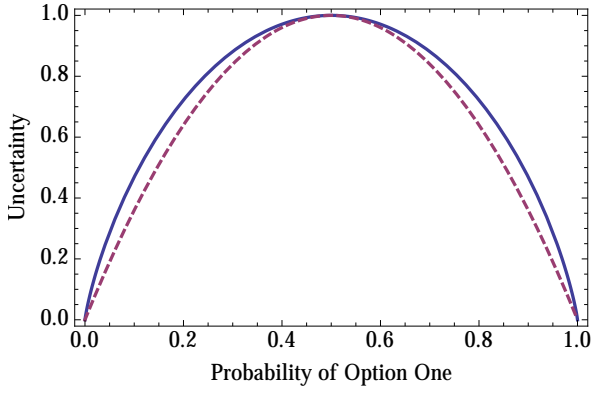


Figure 2: Alternatives to entropy? Blue solid line: the function  $H(X)$  for a binary choice, as a function of the probability of heads. Red dashed line: a common approximation to  $H(X)$  that people sometimes invent, based on the fact that more “uncertain” or “diverse” distributions tend to have lots of small probabilities. They’re close, but not quite overlapping.

first adult, who was biased towards “car”, was your father (he’s an Uber driver). Your mother, by contrast, is a professor of ecology and so she is biased towards trees, choosing “tree”, and not “car”, with probability one-half (and the other two options with probability one-quarter each).

If we use the Dad script on your mother, we get the situation on the right-hand panel of Fig. 3. Even though your mother’s choice has the same uncertainty (and so, therefore, there’s an equally optimal script), the dad script is no longer optimal, and instead of taking 1.5 questions on average, it wastes time by splitting on a less-likely choice, and takes 1.75—an extra quarter-question—on average. That inefficiency can be thought of as due to a failure of expectations, a violation (or refinement) of a prior, or a change in the nature of the information source itself.

It so turns out that this coding failure can also be quantified without explicitly constructing script pairs. It is equal to what is called the Kullback-Leibler divergence (KL), which is defined as

$$KL(p|q) = \sum_{i=1}^N q(x_i) \log_2 \frac{q(x_i)}{p(x_i)}, \quad (6)$$

where  $p(x)$  is the distribution you trained on (built the question script for; your dad), but  $q(x)$  is the new distribution you’re now encountering (the mother). The “Kullback-Leibler from  $p$  to  $q$ ”, or  $KL(p|q)$ , tells you how many additional questions you will ask, on average, over and above what you’d have to ask if you were using an optimal script. (As in the relationship between uncertainty and tree depth, there are caveats about cases where this can be off because the question tree is short; but if one lumps together a bunch of outputs and encodes that, these differences go to zero.)

We can describe KL in terms of question script inefficiencies, or coding failures, and like uncertainty itself, there are many interpretations beyond that. We can talk, for example, about the “surprise” (or “Bayesian surprise”) of one distribution given that you’re expecting another. Fitting in with an epistemic interpretation about agents testing beliefs, It’s also equal to the rate at which evidence in favor of the events being drawn from distribution  $q$  (rather than  $p$ ) accumulate over time—where evidence is defined as the log of the ratio of probability between the  $q$  hypothesis and the  $p$  hypothesis, averaged over lots of trials, and the rate is per sample; see Ref. [2].

For these latter reasons, minimizing the KL of a distribution  $p$  from the “true” distribution  $q$  can be understood as making  $p$  as epistemically close to truth as possible; hence its use in model



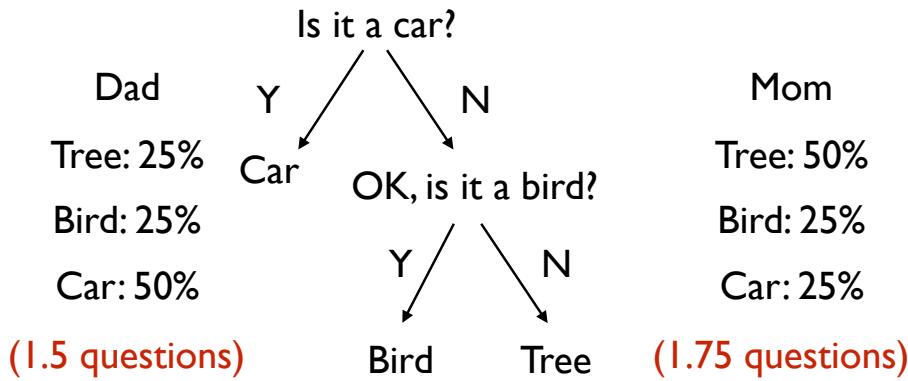


Figure 3: Optimal scripts for twenty questions. Given an opponent (“Dad”) who picks “car” half the time, and the two other options one-quarter of the time each, the game can be solved in an average of 1.5 questions when the script is optimal. But scripts are optimal usually only for the probability distribution they were built for; when an optimal script is applied to a new distribution, it is usually no longer optimal, and inefficiencies in representation—“coding failures”—emerge.

selection, where it is part of the Akaike Information Criterion [3], as well as its role in our new proposal for correcting algorithmic bias when machine learning is used for criminal justice and policy-making [4].

One of the most exciting things I discovered in my work is that the surprise that KL divergence measures has real mental effects. The first use of KL divergence to quantify surprise, or coding failure, in a cognitive system appears to be Itti & Baldi’s work on eye-tracking [5]. They did something very clever: they took a movie clip, and measured the KL divergence spatially over the screen. Formally, they divided up the screen into little patches. In each patch, they computed the probability distribution over pixel colors at time  $t$  (that’s the  $p$  distribution) and then again at time  $t$  plus one second (that’s the  $q$  distribution). Then they could compute the KL divergence between  $p$  and  $q$  in each patch, and (since the movie clip lasts more than a few seconds), actually see how it evolves over time.

Amazingly, their eye-tracking devices showed that people tended to look not (say) at the bright patches, or the uncertain patches, but at the high KL-divergence patches. We tend, in other words, to look towards those parts of a video that violate our expectations (rather than, for example, those parts that just have high uncertainty—we get bored of television fuzz). With some colleagues and students, we went to town on these results, extending them to the cultural and economic domains. Our work on Darwin’s reading patterns [6] is (to our knowledge) the first use of this in the study of natural language (or, indeed, historical archives).

Another paper, in 2018, [7] uses KL in a cultural evolution context, to look at who (and what) is introducing persistent ideas into a system. To do this, we introduced the idea of looking at both KL to the past (“novelty”), and KL to the future (“transience”). We found Robespierre. For a full, and highly excitable list, of why you really want the Kullback-Leibler divergences, see the tweetstorm at <http://bit.ly/kullbackleibler>.

## 7 Einstein and Cromwell’s Rule

People sometimes use KL as a distance measure—“how different are two distributions”—but this is incorrect for a very interesting reason: KL is not necessarily symmetric! The surprise on encountering  $q$  given that you’re expecting  $p$  may not be the surprise on encountering  $p$  given that you’re expecting  $q$ . An organism that grows up in a rich sensory environment will, in general, be less surprised on encountering an improvised one than an organism who has grown up in an impoverished environment and encounters a rich one.

If you’d like an explicit example, consider the difference between me and Einstein. We both sleep and eat a lot, but if you’re watching Einstein he has a small chance  $p$  of inventing a new theory of physics. Say my distribution over sleeping, eating, and inventing new physics is  $\{0.7, 0.3, 0\}$  while Einstein’s distribution is  $\{0.7, 0.2, 0.1\}$ . If you’re expecting Einstein, but you get me, you’re not that surprised; we both spend a lot of time sleeping and eating. But if you expect me, and you get Einstein, you will be very surprised when you see him invent a new theory of physics.

(Because Kullback-Leibler divergence is not symmetric, some people like to invent a distance measure for the “distance” between  $p$  and  $q$  by symmetrizing it:  $KL(p|q) + KL(q|p)$ . This is a suboptimal solution because, while it does make the measure symmetric, it still fails as a measure of distance because it violates the triangle inequality, which requires that the distance from  $p$  to  $q$  must be less than or at worst equal to the distance from  $p$  to  $r$  plus the distance from  $r$  to  $q$ . Below, we’ll see a better metric to use in its place.)

Returning to the Simon-Einstein case, if you think the probability of me inventing a new theory of physics is precisely zero, you will be infinitely surprised—KL divergence will blow up because of the  $0.1 \log(0.1/0)$  term. It’s as if the question script never terminates because the possibility wasn’t written in, or the transmission mechanisms stalls because it encounters a symbol not in its look-up table.

This is a good place to be reminded of how strange it is to attribute zero probability to an event; as you can learn in greater detail in the accompanying article on Bayesian reasoning, if an agent attributes probability zero to an event it means that no evidence whatsoever, no evidence of any form, can convince you that the event took place. This violates what is sometimes referred to as the “Cromwell Rule”, after the 17th Century English military dictator, who famously urged the Church of Scotland “I beseech you, in the bowels of Christ, think it possible that you may be mistaken” (*i.e.*, do not attribute probability zero to something). Practically speaking, if you’re estimating probabilities using a Bayesian method (such as a topic model), you’ll never get a zero value; if you’re using a naive estimator of probabilities that relies on frequencies, you might, and I recommend adding a small regularizing psuedo-count,  $1/k$ , to each bin. See Ref. [8] for more on why this makes sense.<sup>4</sup>

## 8 Mutual Information

Once you can measure  $H(X)$ , the uncertainty in a process, you can talk about how that uncertainty changes when you acquire information from somewhere else. For example, while you might be

---

<sup>4</sup>The Simon-Einstein case also provides a nice example of how KL is an asymptotic measure of inefficiency. If my probability of revolutionizing physics is non-zero (call it  $\epsilon$ ), then because an observer needs to code for all three possibilities, they will end up using a Y/NY/NN code for both of us, with Y coding sleep, and the inefficiency will be zero even though KL grows as  $\epsilon$  becomes small. It’s only when we start encoding multiple moments at once that the differences in coding efficiency per event approach the KL value—the Einstein code will start to make efficient codes for combinations of moments that include revolutionizing physics well before the Simon code will.

maximally uncertain about the card on the top of a shuffled deck, your uncertainty goes down a great deal once I tell you the suit. Instead of having an equal choice over 52 cards (an entropy of  $\log_2 52$ , about 5.7 bits), you have an equal choice over the thirteen cards of that suit ( $\log_2 13$ , or 3.7, bits). Put another way, telling you the suit was worth two bits of information.

In general, if two variables,  $X$  and  $Y$ , are correlated with each other, this means that the probability distribution over one of the variables is affected by the value of the other variable. If it's raining, then I'm more likely to be carrying an umbrella; more formally,  $P(X|y)$  will be different depending on the value of  $y$ , and the uncertainty of the distribution over  $X$  given that  $Y$  is equal to  $y$ , which we write as  $H(X|y)$ , will differ from the uncertainty over  $X$  before you got this information,  $H(X)$ .

If we average over all the possible values of  $Y$ , weighted by their probability, we get what is called the conditional entropy of  $X$  given  $Y$ , or  $H(X|Y)$ ,

$$H(X|Y) = \sum_{j \in M} H(X|y_j)P(y_j) \quad (7)$$

where we assume that there are  $M$  possible values of  $Y$ . In some cases,  $H(X|y)$  can be larger than  $H(X)$ ; if, for example,  $X$  gets more random under unusual conditions. However, it can be shown (using something called Jensen's inequality) that on average, *information never hurts*. In other words,  $H(X|Y)$  is always less than, or (in the case that  $Y$  and  $X$  are independent) equal to  $H(X)$ .

We usually talk in terms of the extent to which  $Y$  reduces uncertainty about  $X$ ; the mutual information,  $I$ , between  $X$  and  $Y$  is thus defined as

$$I(X, Y) = H(X) - H(X|Y), \quad (8)$$

and a little algebra suffices to show that, in fact, this is symmetric—the information that  $Y$  gives you about  $X$  is equal to the information that  $X$  gives you about  $Y$ ; or  $H(X) - H(X|Y)$  is equal to  $H(Y) - H(Y|X)$ . How could it not be?—if you tell me something and I learn about the world, I can reverse the process, look at the world, and figure out what you'll say about it.

Mutual Information is at the heart of a number of problems where people try to figure out how much information, or influence, is propagating between systems, from one part of the country to another, all the way down to one word to the next in a text. You can think of it as the “most general” form of correlation coefficient that you can measure. In an analogous fashion to how uncertainty talks about the optimal question script (without needing to specify it in details), mutual information talks about the optimal method for learning about one thing using information about another.

## 9 Jensen-Shannon Distance

Let's return to Mom, Dad, and twenty questions one last time. Say you're playing over a computer with one of them, but you don't know which one. Over time, as you play, you will start to accumulate information in favor of one or the other—you'll notice a preponderance of “car” choices, indicating your father, or “tree” choices, indicating your mother.

How much information about the identity of the player do you get from playing a single round? From the symmetry of mutual of information, this is equivalent to how much information about the identity of the word you get from learning who the player is, or how much your uncertainty is reduced, on average, by gaining this. This is called the Jensen-Shannon Distance, or JSD, between  $P$  and  $Q$ :

$$JSD(P, Q) = H(M) - \frac{1}{2}(H(P) + H(Q)), \quad (9)$$

where  $M$  is the mixture distribution of the two adults,  $m_i$  is equal to  $\frac{1}{2}(p_i + q_i)$ . In words, the JSD tells you how much one sample, on average serves to distinguish between the two possibilities. Once you know which distribution you’re drawing from you will (on average) be able to construct a more efficient question tree. The unit of JSD is (of course) bits—the number of yes/no questions you get for free (again, on average) once you know which of the two distributions,  $P$  or  $Q$ , you’ll be drawing words from.

A simple computation shows that this is equal to the KL from  $M$  to  $P$  plus the KL from  $M$  to  $Q$ ,

$$JSD(P, Q) = \frac{1}{2}(KL(M|P) + KL(M|Q)) \quad (10)$$

A very nice feature of the JSD is that it (or, rather its square-root) is not just symmetric—it’s also a metric that obeys the triangle inequality [9], making it a good tool for using dimensionality reduction tools such as multidimensional scaling (MDS, such as the `cmdscale` in R) that work best with sensible distances. Because  $m_i$  is zero if and only if both  $p_i$  and  $q_i$  are zero, it’s also the case that JSD never blows up.

For these reasons, JSD serves well as a method for understanding distinctiveness a “distance” between two probability distributions. It plays the same role for probabilities as Euclidean distance does for ordinary “spatial” vectors. Ref. [10] was this author’s introduction to its use in empirical work, where it quantifies the distinguishability of two “genres” of court case in the 18th and 19th Centuries.

## 10 A Note on Measuring Information

Talking about information, and proving theorems about it, is one thing. Measuring the creation, flow, and extinction of information in the real world, on data gathered in the field, is a separate task.

In many cases, the “naive”, or “plug-in” approach to measuring information-theoretic quantities in the real world works very well. You begin by estimating the probability of events from frequency counts; given these estimated probabilities, it’s a simple matter to plug them in to the formula you care about—whether it be uncertainty, Kullback-Leibler divergence, mutual information, or Jensen-Shannon Distance.

If you have enough data, and not too many categories, this will usually work very well. A good rule of thumb is that you should have at least ten times as many samples as you do classes or categories that you’re measuring [11]. So, twenty tosses of a coin are sufficient to get a good estimate of the uncertainty of the distribution; forty samples of two stocks will be enough to determine the mutual information between their “up” or “down” ticks (a 2x2 grid of possibilities—stock one up while stock two down, and so on).

When you don’t have enough data—or you want to do more sophisticated things, like determine error bars on estimates—you have to be a bit smarter. The estimation of information theoretic quantities from data has a big history (often independently rediscovered in different fields). Ref. [11] describes some of the issues that arise, and THOTH (<http://thoth-python.org>) implements some of the most common tools in python. When you are measuring the uncertainty of distributions over very large cardinalities, consider the NSB estimator [8] <http://nsb-entropy.sourceforge.net>; over continuous spaces, the NPEET toolkit <http://www.isi.edu/~gregv/npeet.html> which implements a few [12, 13, 14].

## 11 Minds and Information

Information theory is fundamentally about signals, not the meaning they carry. What we measure thus requires interpretation; that I am uncertain about different options may not be as important as the fact that I'm uncertain about these particular two possibilities. Information theory tells you what you know, but it doesn't tell you what matters; you need something like a utility function from the decision theorists or game theories to tell you about that. Even though it is about signals and signal transmission, it can't tell you how those signals emerged; you'll need to supplement a story about agents exchanging information with, for example, a Brian Skyrms-like story about the emergence of signaling systems [15].

Conversely, in its universality, information theory applies just as much to the written and spoken words of humans as to the electronic machines for which it was first developed. And it allows us to compare distant worlds—no more, and no less, exciting than, say, comparing the real income of an English bricklayer in 1350 to one in 1780, the hours worked by a French housewife in 1810 and 1950, or the life expectancy of a hunter-gatherer of the Maasai to that of a child in a Manchester factory of 1840.

That we can *quantify* information is both intriguing and mysterious. Intriguing, because information is one of the fundamental features of our minds and our social worlds. History, psychology, economics, cognitive science, economics—all would grind to a halt were their practitioners forbidden from using the concept of information at will. Mysterious, because information is a fundamentally epistemic property: it is about what one knows, and is, as such, relative to that observer in a way that one's (real or nominal) salary, height, daily caloric intake, or place and date of birth are not.

Subject-relative facts, of course, abound—facts about trust, say, or allegiance, virtue, belief, love—and they make up a core part of the worlds we want to understand. What we learned in the twentieth century is that at least one such fact, the information one has, *can* be quantified. The economists have tried to quantify another set of subject-relative facts, one's desires, through utility theory, with somewhat less empirical success. Facts on the edge between reality and perception include those concerning inequality, and we have made a great deal of progress in figuring out both how to measure inequality, and what its implications are [16].

Many years ago, Shannon—the originator of information theory—wrote a somewhat dispirited article titled “the Bandwagon” [17] that worried about the indiscriminate use of information theory in other fields. In the modern era, it appears that Shannon's objections have, by and large, been answered.

In particular, the idealized nature of information theoretic concepts is now well understood. We realize that an information theoretic quantity provides a limit—sometimes, but not always reached—on how well a real system can perform.

Meanwhile, progress towards testing the extent to which information theoretic quantities do, and do not, apply to real cognitive and social phenomena, though at first slow in coming, is now beginning to build. New, often very large, data-sets give us the sample sizes we need to test sophisticated hypotheses about what underlying mechanisms might be tracked by quantities such as KL, JSD, and Mutual Information. Meanwhile, the Bayesian turn in cognitive science [18, 19] has found new uses for the optimal descriptions of reasoning provided by probability theory and Bayes' rule, providing new theoretical backing for the use of information theory to quantify the resulting distributions.

Taken together, it seems clear that, while the study of information in the social world is in its infancy, but not without some recent successes under its belt. This author has worked on a few of them, and expects to see more in the cognitive science to come.

## References

- [1] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.
- [2] Jonathon Shlens. Notes on Kullback-Leibler divergence and likelihood. *arXiv preprint*, arXiv:1404.2000, 2014. <https://arxiv.org/abs/1512.04177>.
- [3] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.
- [4] Simon DeDeo. Wrong side of the tracks: Big data and protected categories. In Cassidy R. Sugimoto, Hamid Ekbia, and Michael Mattioli, editors, *Big Data is Not a Monolith*. MIT Press, Cambridge, MA, USA, 2016. arXiv:1412.4643; <https://arxiv.org/abs/1412.4643>.
- [5] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306, 2009.
- [6] Jaimie Murdock, Colin Allen, and Simon DeDeo. Exploration and exploitation of victorian science in darwin’s reading notebooks. *arXiv preprint*, arXiv:1509.07175, 2015. <https://arxiv.org/abs/1509.07175>.
- [7] Alexander TJ Barron, Jenny Huang, Rebecca L Spang, and Simon DeDeo. Individuals, institutions, and innovation in the debates of the french revolution. *Proceedings of the National Academy of Sciences*, 115(18):4607–4612, 2018.
- [8] Ilya Nemenman, William Bialek, and Rob de Ruyter van Steveninck. Entropy and information in neural spike trains: Progress on the sampling problem. *Physical Review E*, 69(5):056111, 2004.
- [9] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, Jan 1991.
- [10] Sara Klingenstein, Tim Hitchcock, and Simon DeDeo. The civilizing process in London’s Old Bailey. *Proceedings of the National Academy of Sciences*, 111(26):9419–9424, 2014.
- [11] Simon DeDeo, Robert X. D. Hawkins, Sara Klingenstein, and Tim Hitchcock. Bootstrap methods for the empirical study of decision-making and information flows in social systems. *Entropy*, 15(6):2246–2276, 2013.
- [12] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, Jun 2004.
- [13] Greg Ver Steeg and Aram Galstyan. Information transfer in social media. *CoRR*, abs/1110.2724, 2011. <http://arxiv.org/abs/1110.2724>.
- [14] Greg Ver Steeg and Aram Galstyan. Inferring predictive links in social media using content transfer. *CoRR*, abs/1208.4475, 2012. <http://arxiv.org/abs/1208.4475>.
- [15] Brian Skyrms. *Signals: Evolution, learning, and information*. Oxford University Press, Oxford, UK, 2010.

- [16] Thomas Piketty. *Capital in the twenty-first century*. Belknap Press, 2014. Translated by Arthur Goldhammer.
- [17] C. Shannon. The bandwagon. *IRE Transactions on Information Theory*, 2(1):3–3, March 1956.
- [18] Thomas L. Griffiths, Charles Kemp, and Joshua B. Tenenbaum. Bayesian models of cognition. In Ron Sun, editor, *The Cambridge Handbook of Computational Psychology*. Cambridge University Press, Cambridge, UK, 2008.
- [19] Nick Chater, Mike Oaksford, Ulrike Hahn, and Evan Heit. Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(6):811–823, 2010.

# Bayesian Reasoning for Intelligent People

Simon DeDeo\*

November 10, 2021

## Contents

|           |  |           |
|-----------|--|-----------|
| <b>1</b>  | <b>The Bayesian Angel</b>  | <b>1</b>  |
| <b>2</b>  | <b>Bayes' Theorem and Madame Blavatsky</b>                       | <b>3</b>  |
| <b>3</b>  | <b>Observer Reliability and Hume's Argument against Miracles</b> | <b>4</b>  |
| <b>4</b>  | <b>John Maynard Keynes and Putting Numbers into Minds</b>        | <b>7</b>  |
| <b>5</b>  | <b>Neutrinos, Cable News, and Aumann's Agreement Theorem</b>     | <b>10</b> |
| <b>6</b>  | <b>Specifying Priors and the Zen Koan of Marvin Minsky</b>       | <b>13</b> |
| <b>7</b>  | <b>Further Reading</b>   | <b>15</b> |
| <b>8</b>  | <b>Technical Notes on Working with Probabilities</b>             | <b>16</b> |
| <b>9</b>  | <b>Nate Silver, Sam Wang, and who will win the election?</b>     | <b>18</b> |
| <b>10</b> | <b>Waiting for the Number 55 bus</b>                             | <b>23</b> |
| <b>11</b> | <b>Acknowledgements</b>  | <b>27</b> |

## 1 The Bayesian Angel

“Bayesian reasoning” is a fancy phrase for “the use of probabilities to represent degrees of belief, and the manipulation of those probabilities in accordance with the standard rules.” You learned many of the standard rules for manipulating probability in high school; you can find a derivation of them in Ref. [1] (Lecture Four, “Laplace’s Model of Common Sense”). There are, in fact, many ways to derive them, ranging from the philosophical (“consistent reason requires us to do this”), to the evolutionary (“populations whose members approximate these rules better, grow more quickly”),

---

\*Social and Decision Sciences, Carnegie Mellon University & the Santa Fe Institute. Current version available at <http://santafe.edu/~simon/br.pdf>. Please send corrections, comments and feedback to [simon@santafe.edu](mailto:simon@santafe.edu); <http://santafe.edu/~simon>.



to the economic (“if you don’t do this, I can induce you to place a series of bets with me that you are guaranteed to lose.”)

The most basic of these rules is how to turn joint probabilities into conditional probabilities. We have the following identity:

$$P(x, y) = P(x|y)P(y) = P(y|x)P(x), \quad (1)$$

where  $P(x, y)$  is notation for “the probability of both  $x$  and  $y$  being true” (the “joint” probability),  $P(x|y)$  is notation for “the probability of  $x$  being true *conditional* on (or given that)  $y$  being true”, and  $P(x)$  is notation for “the probability that  $x$  is true”. Here  $x$  and  $y$  are shorthand for sentences; for example,  $x$  might stand for “the student chosen at random is blond”, and  $y$  might stand for “the student chosen at random is male.” You should try saying Eq. 1 out loud, using explicit sentences for  $x$  and  $y$ , to see if it sounds reasonable to you.

That’s it. Just thinking carefully about how to set up problems, and using this equation at critical moments, will make you an ideal reasoner—in an important sense, an optimal reasoner.<sup>1</sup> Using Bayes’ theorem might make you an optimal reasoner, but it’s not something an ordinary person can do all the time, at least, not perfectly. It might, for example, require keeping track of gigantic lists of conditional probabilities and manipulating them at will, and that’s not something we evolved to do well. Sometimes, to emphasize the idealized nature of Bayesian descriptions of reasoning, and to contrast them with the messier kind of approximations (or forgeries) that human beings do, we’ll talk about a *Bayesian angel* with infinite memory and processing power (but, importantly, not infinite perceptual abilities nor experience—like bodhisattvas, Bayesian angels live in the real world).

While Bayesian reasoning makes no changes to how you follow the rules you learned in high school, it does ask you to make a fundamental shift in how you think about them. You are likely used to thinking about probabilities in terms of frequencies: if the probability of an event  $x$  is 0.5, you expect it to happen “about 50% of the time”. We call this the *frequentist* perspective, though it comes so naturally to us we can forget it’s a deliberate choice. Often times the frequentist perspective is good enough: if you have a coin, you can toss it multiple times to see what happens and use the number of times it comes up heads to attribute an probability to the coin itself. Frequencies are facts about the world, and so the frequentist can be said to think of probabilities as themselves “objective” facts about the world.

Bayesians flip the problem around: instead of seeing probabilities as out there in the world, a property of objects like coins, they understand them as describing subjective states of belief that an observer might have. If an individual attributes a probability  $p$  to event  $x$ , this is now understood as indicating that the observer has “degree of belief  $p$ ” in the event’s taking place (or, indeed, in having taken place). If  $p$  is close to one, the individual is very sure  $x$  is true (or will happen); if  $p$  is close to zero, the individual is very sure it’s not the case. If  $p$  is precisely zero, the individual considers it absolutely impossible; as we shall see, this means that (for example) no evidence of any form whatsoever will cause them to raise  $p$  above zero.

The shift from “frequentist” to “subjective” probability is not an easy one to make. But it’s absolutely essential to do it if you want to understand the Bayesian revolution in cognitive science, where it plays a core role in modeling actual states of belief in real-world agents—or, indeed, if you want the ways in which these ideas have revolutionized artificial intelligence, machine learning, and data analysis in the modern era.

---

<sup>1</sup>You can be an optimal reasoner with the wrong facts, or an optimal reasoner with the right facts and the wrong belief space (set of sentences). We’ll see examples of both later in this guide. Being optimal doesn’t mean being *right* about anything.

The fact that Bayesians understand probabilities as describing a subjective, reasoning process may seem to undermine the “objectivity” of the statements it makes. In fact, it turns out to be a huge advantage: think, for example, of predicting the outcome of an election. If we’re told our candidate has a 70% chance of winning, we want to know why—on what basis—was that prediction made? We want to see how the degree of belief the statistician urges on us is a combination of other things we might believe, given polling data, say, and theories, which we may only have limited confidence in, about how to interpret it. Bayesian tools lift the cover on this process, laying the machinery of thought bare for inspection.

## 2 Bayes’ Theorem and Madame Blavatsky

The identity Eq. 1 is the basis of something famously known as “Bayes’ Theorem”. You just divide both sides by  $P(y)$ :

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} \quad (2)$$

Bayes’ Theorem is something you can use to win arguments against statistics nerds, and telling them they’ve violated it is sort of like telling an ordinary person that they have a rip in the seat of their pants. They are going to want to check that out right away. Bayes’ Theorem gets its power from how it can invert a kind of question you tend to ask, but can’t answer, into a question that sounds a bit weird, but it turns out you can answer. We’ll use  $T$  and  $D$  as our random variables now, where  $T$  stands for a set of (exhaustive and mutually exclusive) theories about the world, and  $D$  for the kind of data you might get.

We often want to know what our degree of belief should be in a theory  $t$ , given some data  $d$ . This is a profound problem when the theories are of great import and the data expensive and hard to gather. To have some fun, we’ll do some toy problems: made-up scenarios that will help you build your intuition and give you the skills to reason when it matters. We begin with an example drawn from the physicist E.T. Jaynes, in the “Queer Uses of Probability” chapter of Ref. [1].

Our friend Artemy<sup>2</sup> returns from a trip to New York City, reporting that he saw Madame Blavatsky, the famous clairvoyant, successfully predict the outcome of 100 coin tosses. Should we believe in ESP, the theory that some people have a magical ability to sense the future?

We start by setting up our  $T$  and  $D$ .  $T$  is {“ESP is real”, “ESP is not real”}.  $D$  is {“Madame Blavatsky is no better than chance at predicting the toss of an unbiased coin”, “Madame Blavatsky can predict perfectly the outcome of 100 coin tosses”} (for simplicity, we’ll assume these are the only possible kinds of data we can get—another way to put it is that we’ll attribute probability zero to any other kind of data). We’ll abbreviate these as  $T=\{\text{ESP}, \overline{\text{ESP}}\}$  and  $D=\{\text{normal}, \text{predict}\}$ ; the little overline is a standard way to notate “not”.

We want to know the following: given that Madame Blavatsky did this amazing thing, what should I believe about ESP? More formally, “conditional on *predict*, what degree of belief should I have in *ESP*?” Using a rearrangement of Eq. 1, we have

$$P(\text{ESP}|\text{predict}) = \frac{P(\text{predict}|\text{ESP})P(\text{ESP})}{P(\text{predict})}. \quad (3)$$

We’ll go term by term on the right-hand side.  $P(\text{predict}|\text{ESP})$  means “what’s the chance that we get the data {Madame Blavatsky predicts perfectly} given the truth of the theory ESP.” Let’s say

---

<sup>2</sup>A parallel universe version of one-time fearless Associate Instructor at IU, now postdoctoral fellow at the Santa Fe Institute, Artemy Kolchinsky.

that if ESP is real, Madame Blavatsky almost certainly has it, and if she has it, she can do amazing predictions like these, so we set that at 0.9—*i.e.*, only a 10% chance she’ll screw up using her (real) magic powers.

$P(\text{ESP})$  is the prior belief you have in ESP—the degree of belief you attribute to the possibility before hearing about the new data. Let’s say you’re a scientist; you attribute low value to these kinds of things, but (you’re a scientist)—nothing is impossible, so we’ll say  $10^{-12}$ . You’re more confident that ESP is fake than you are about surviving your next airline flight.<sup>3</sup>

Finally,  $P(\text{predict})$ : the probability this prediction event happens. I always find this term hard to think about, but then I just recall that  $P(\text{ESP}|\text{predict}) + P(\overline{\text{ESP}}|\text{predict})$  has to sum to unity.<sup>4</sup>

$$P(\text{predict}) = P(\text{predict}|\text{ESP})P(\text{ESP}) + P(\text{predict}|\overline{\text{ESP}})P(\overline{\text{ESP}}).$$

$P(\overline{\text{ESP}})$  is easy—that’s just  $1 - P(\text{ESP})$ , or  $1 - 10^{-12}$ .  $P(\text{predict}|\overline{\text{ESP}})$  is the chance of guessing one hundred coin tosses in a row, given the fact that it’s impossible to see the future (so you have to guess). That is just  $0.5 \times 0.5 \times \dots$ —you have a fifty-fifty chance the first time, times a fifty-fifty chance the second, and so forth. For 100 tosses, it’s  $2^{-100}$ , or about  $7 \times 10^{-31}$ . We’re set—we can now plug in all the numbers to discover that

$$P(\text{ESP}|\text{predict}) = \frac{0.9 \times 10^{-12}}{0.9 \times 10^{-12} + 7 \times 10^{-31}(1 - 10^{-12})} \approx 1 - 10^{-18}, \quad (4)$$

or, in words: ESP is almost certainly true (very very very close to one), conditional on Madame Blavatsky’s performance.

### 3 Observer Reliability and Hume’s Argument against Miracles

Or is it? Let’s go back to our theory set,  $T$ , and enlarge it. What if we allowed for another possibility: our friend Artemy is delusional, or was gulled by a stage magician?  $T$  is now {“ESP is real, Artemy is not crazy”, “ESP is not real, Artemy is not crazy”, “ESP is real, Artemy is crazy”, “ESP is not real, Artemy is crazy”}.

If we label the belief that “Artemy is crazy” as AC, we can then abbreviate these four options, symbolically, as  $\{\text{ESP}+\overline{\text{AC}}, \overline{\text{ESP}}+\overline{\text{AC}}, \text{ESP}+\text{AC}, \overline{\text{ESP}}+\text{AC}\}$ . We can assume, for simplicity, that in the prior, these things appear independently, so that

$$P(\text{ESP}+\overline{\text{AC}}) = P(\text{ESP})P(\overline{\text{AC}}). \quad (5)$$

---

<sup>3</sup>You might say “it’s more likely I’ll die in a plane crash on my next flight than ESP turns out to be real.” True Bayesians would object to this statement, because I’m implicitly referring to a frequentist notion—“what *would* happen *if* you tried something”; even worse, implicitly, “how often would it happen if you tried it lots of times”. Frequentists are the ancient and heretical pagans of probability, whose practices were superseded by the unitary Bayesian religion. To be admitted to the Church of Bayes, you must remain entirely in the world of degrees of belief. Note that in this case, I (Simon) *used* counts of airline crashes to derive a degree of belief—but that’s OK, and I’m still a Bayesian, because secretly I did something a bit meta: I used the counts to test different hypotheses about the degree of belief I should have in airline safety, and chose the degree of belief about airline safety that I had the highest degree of belief in! I used a Dirichlet prior with  $\alpha = 1$  (now do you believe me). Bayesians are allowed to have data commerce with Frequentists. They just can’t share beliefs—because, a True Bayesian would say, Frequentists are fundamentally mistaken about the nature of beliefs they have, and so can’t communicate them to you. Frequentists will, for example, eventually give contradictory answers to logically identical questions.

<sup>4</sup>Rewrite Eq. 3 for  $P(\sim \text{ESP}|\text{predict})$ , set the sum of that and Eq. 3 equal to 1, and solve for  $P(\text{predict})$ .

If the prior didn't decompose, it would mean that somehow these two features of the world would be connected. In an ESP world, for example, Artemy might be more likely to be crazy.

We consider Artemy a usually very reliable guy, so the chance that he's crazy, let's say, is  $10^{-6}$ . It's very unlikely that Artemy is crazy—about ten times less than the lifetime chance of dying from being struck by lightning.<sup>5</sup> Now,

$$P(\text{ESP} + \overline{\text{AC}} | \text{predict}) = \frac{P(\text{predict} | \text{ESP} + \overline{\text{AC}})P(\text{ESP})P(\overline{\text{AC}})}{P(\text{predict})}. \quad (6)$$

Let's say that if Artemy is not crazy, but ESP is real,  $P(\text{predict} | \text{ESP} + \overline{\text{AC}})$  is just  $P(\text{predict} | \text{ESP})$ —we previously agreed that that was 0.9. But now the denominator has changed—we'll compute  $P(\text{predict})$  by summing over four possibilities, not two. Rather than be tedious about it, let's ask a slightly different question. What's the *odds-ratio* of  $\overline{\text{ESP}} + \text{AC}$  vs.  $\text{ESP} + \overline{\text{AC}}$ ? How much more likely is it that ESP is false, and Artemy is crazy, rather than ESP is true and Artemy not crazy? We'll just divide the two,

$$\frac{P(\overline{\text{ESP}} + \text{AC} | \text{predict})}{P(\text{ESP} + \overline{\text{AC}} | \text{predict})} = \frac{P(\text{predict} | \overline{\text{ESP}} + \text{AC})P(\overline{\text{ESP}})P(\text{AC})}{P(\text{predict} | \text{ESP} + \overline{\text{AC}})P(\text{ESP})P(\overline{\text{AC}})}, \quad (7)$$

where (to be clear) we know the denominator from Eq. 6, we had to use Eq. 1 to get the numerator, and that annoying  $P(\text{predict})$  cancelled. We just need to specify our theory of what happens when ESP is false, but Artemy is crazy. In this case, let's say, Artemy can be led to believe crazy things by a pseudomagician without much difficulty, and  $P(\text{predict} | \overline{\text{ESP}} + \text{AC})$  is, let's say, 0.9.

$$\frac{P(\overline{\text{ESP}} + \text{AC} | \text{predict})}{P(\text{ESP} + \overline{\text{AC}} | \text{predict})} = \frac{0.9 \times (1 - 10^{-12}) \times 10^{-6}}{0.9 \times 10^{-12} \times (1 - 10^{-6})} \approx 10^6, \quad (8)$$

or, in words: it's a million times more likely that Artemy is crazy, than ESP is real.

This mathematical result was anticipated well before Bayes' rule gained the fame it has today. The famous free-thinker, philosopher, and Scotsman, David Hume wrote in his essay "Of Miracles" in 1784,

The plain consequence is (and it is a general maxim worthy of our attention), "That no testimony is sufficient to establish a miracle, unless the testimony be of such a kind, that its falsehood would be more miraculous, than the fact, which it endeavours to establish: And even in that case, there is a mutual destruction of arguments, and the superior only gives us an assurance suitable to that degree of force, which remains, after deducting the inferior."

When anyone tells me, that he saw a dead man restored to life, I immediately consider with myself, whether it be more probable, that this person should either deceive or be deceived, or that the fact, which he relates, should really have happened. I weigh the one miracle against the other; and according to the superiority, which I discover, I pronounce my decision, and always reject the greater miracle. If the falsehood of his testimony would be more miraculous, than the event which he relates; then, and not till then, can he pretend to command my belief or opinion.

---

<sup>5</sup>National Safety Council, *Injury Facts 2013*. [http://www.nsc.org/news\\_resources/injury\\_and\\_death\\_statistics/Documents/Injury\\_Facts\\_43.pdf](http://www.nsc.org/news_resources/injury_and_death_statistics/Documents/Injury_Facts_43.pdf); the reader is referred to footnote 6.

Notice one last, sad, thing about Artemy and ESP. Say Artemy tells me about the ESP he saw, and my priors mean that I think he’s crazy. He gets upset—and comes back the next day. “You didn’t believe me yesterday, Simon, but I went back and *I saw her do it all over again*. Now will you believe me?”

The major effect the new information has is to raise the stakes: it makes  $P(\overline{\text{ESP}} + \overline{\text{AC}}|\text{predict})$ , the possibility that everything is normal in the world, even smaller. Let’s take the “normie” perspective, and first look at the odds ratio between “everything is normal, ESP isn’t real, my TA is fine” and “everything is normal, ESP isn’t real, except oh God my TA just had a psychotic break”:

$$\frac{P(\overline{\text{ESP}} + \overline{\text{AC}}|\text{predict})}{P(\overline{\text{ESP}} + \text{AC}|\text{predict})} = \frac{2^{-n} \times (1 - 10^{-12}) \times (1 - 10^{-6})}{0.9 \times (1 - 10^{-12}) \times 10^{-6}}, \quad (9)$$

where  $n$  is the number of correct coin tosses that Artemy claims to have seen. This ratio can be built as with all the others, by going back (a little tediously) to Bayes’ rule. The numerator of the fraction includes the probability of getting the result by (honest) chance, along with the very strong priors in favor of ESP not being real and Artemy being sane. The denominator includes the factors associated with how easy it would be to trick a (crazy) Artemy, the strong normie prior against ESP, and the strong prior against Artemy being crazy.

As  $n$  gets larger, the numerator gets smaller and smaller. When  $n$  is small (say five), the odds are in favor of a lucky outcome for the coin tosses—Artemy wasn’t tricked, it was just a coincidence. But when  $n$  gets larger than about twenty, the  $2^{-n}$  factor gets smaller than the prior against Artemy being crazy, and now, more and more, the data coming in suggest that, in fact, Artemy is just crazy.

In other words, not only does this new evidence not lead me to believe in ESP—it actually increases my confidence that he’s crazy! He provides me what he thinks is more and more evidence for ESP, but my priors mean I take it as evidence for something else. It is not just that “extraordinary claims require extraordinary evidence”—they also take extraordinary faith in the reliability of the evidence-provider.

There are, in the end, four possible things that we are choosing between. Let’s call them:

1. Normie world 1. ESP isn’t real, Artemy is fine.
2. Normie world 2. ESP isn’t real, but Artemy is crazy and seeing things.
3. Enchanted world. ESP is real, and Artemy is a visionary.
4. Comedy world. ESP is real, but Artemy is crazy and seeing things.

As evidence comes in, we shift away from option one into (mostly) options two and three. (Option four, which reminds me of what the philosophers call Gettier cases, is really hard to believe in, because it requires you to believe two impossible things.) Unfortunately for Artemy, *which* of the two options we shift into depends on our priors. If we have normie priors (where ESP is less likely than our friend being insane), our friend’s insanity looks like, well, insanity and we shift belief from Normie World 1 to Normie World 2. But if (exercise left to the reader) our priors against ESP were weaker, and our priors in favor of our friend stronger, we would shift to Enchanted World.

We’ve taken an extreme example here, with some things millions of times more likely than others. The basic pattern, however, where people interpret the same evidence and draw opposite conclusions, is common. We’ll return to this in Sec. 5 below.

## 4 John Maynard Keynes and Putting Numbers into Minds

Ten percent; zero point nine; ten to the negative eighteen. To be fluent in Bayes means to be happy putting numbers onto degrees of belief. But where do they come from? We don't make *all* of them up: for example, we derived the tiny degree of belief  $P(\text{predict} | \sim \text{ESP})$  to be  $7 \times 10^{-31}$  from saying that the probability of an unbiased (unmagical) coin landing heads was 50%. Meanwhile, the very point of Bayes' theorem is that you get more out than you put in;  $P(\text{predict} | \text{ESP})$  is something we derived.

Yet we did have to put some things in, like my prior degree of belief that Artemy is crazy. Where do these numbers come from? If you look in the footnotes, I jokingly relate the degree of belief I have in Artemy being crazy to the likelihood I attribute to getting struck by lightning. And perhaps we do learn something about the strength of beliefs like this, by introspection—this is just about as likely as that; this is less likely; this is more. We spend a lot of time asking ourselves what we believe more, and while we have all sorts of ways of doing that (deductive reasoning, guessing, gut feel) a lot of these activities cash out as comparing and ranking. We can compare and rank numbers, too (“is  $x$  bigger than  $y$ ?”). So perhaps it's not crazy to attach numbers to beliefs, in ways that respect our ranking and ordering, to keep track of exactly what's more likely than what.

But it's not everything, and in particular it's not quantifying. How can we anchor these beliefs—say not just that A is more likely than B ( $P(A) > P(B)$ ), but that  $P(A)$  is, say, 0.1? A classic route to putting numbers on beliefs is via the gambling table. I measure your degree of belief in something by the amount you're willing to wager on it, and the odds you're willing to take. Such a move has been made since the eighteenth century.<sup>6</sup>

You're really sure that something is true? You attach degree of belief 99% to it? Well, if you think it's only 1% likely to be false, then you'll be happy to take a bet where you get \$1 if it's

---

<sup>6</sup>Why not before? For a long time, it was remarkable to me that the connection between probability and gambling took so long to be made. The Romans gambled, but didn't compute probabilities? Newton invented the calculus before we knew how to wager on blackjack?

Was it just that, for some philosophical or theological reason having to do with an all-seeing God, humans could not imagine chance before the eighteenth century? Impossible, of course: if anything, they were more acquainted with the *rota fortunae*, the wheel of fortune, than later centuries. So the failure to mathematize probability seemed crazy to me, because a little such knowledge gives you a huge advantage in games of chance. It was as if the human race conspired to leave money on the table. It all seemed crazy, that is, until I looked at the ways in which people actually played games, and looked at the material objects they played them with.

If you do this, you learn that the dice that Roman centurions must have tossed at the foot of the cross were not the machined tools of twentieth century Las Vegas; and the backs of the cards in Carravaggio's *Cardsharps* a far cry from the mechanically reproduced and indistinguishable patterns that anonymize them today. All games of chance, I believe, were biased in unexpected ways by the objects one played them with. One built up knowledge of the biases in the dice, say, over time, and wagered and bet dynamically as one watched others at the game. The tools of gaming were *themselves* subject to both learning (how biased is this die) and negotiation (if the biases can be manipulated, can I maneuver the game in such a way that I get to do it?) Far too complex for probability theory which deals, at least in its most elementary form, with stationary cases where probabilities are fixed, and known, ahead of time. Such conditions did not emerge until the tools to machine such systems were invented. I haven't proven this theory yet, but it's a good one.

*Update*, from Simon Crase of the Nelson Science Society, New Zealand: Girolamo Cardano's *Liber de Ludo Aleae* was written a century before the *Principia*. According to Leonard Mlodinow, Cardano paid his way through medical school by gambling, exploiting his invention of probability theory. I imagine that he weighed the value of writing his book versus having a nice little earner to fall back on. In Cardano's day there was a lot of status attached to possessing a secret, such as Tartaglia's solution to the cubic; by Newton's day status came from publishing, showing that you could do things that Hooke couldn't. Maybe the real difference came about because of the Royal Society and Sir Christopher Wren. [My response: I'd like to see a historian of science, rather than a physicist, establish that Cardano did indeed use “probability theory” for gambling!]

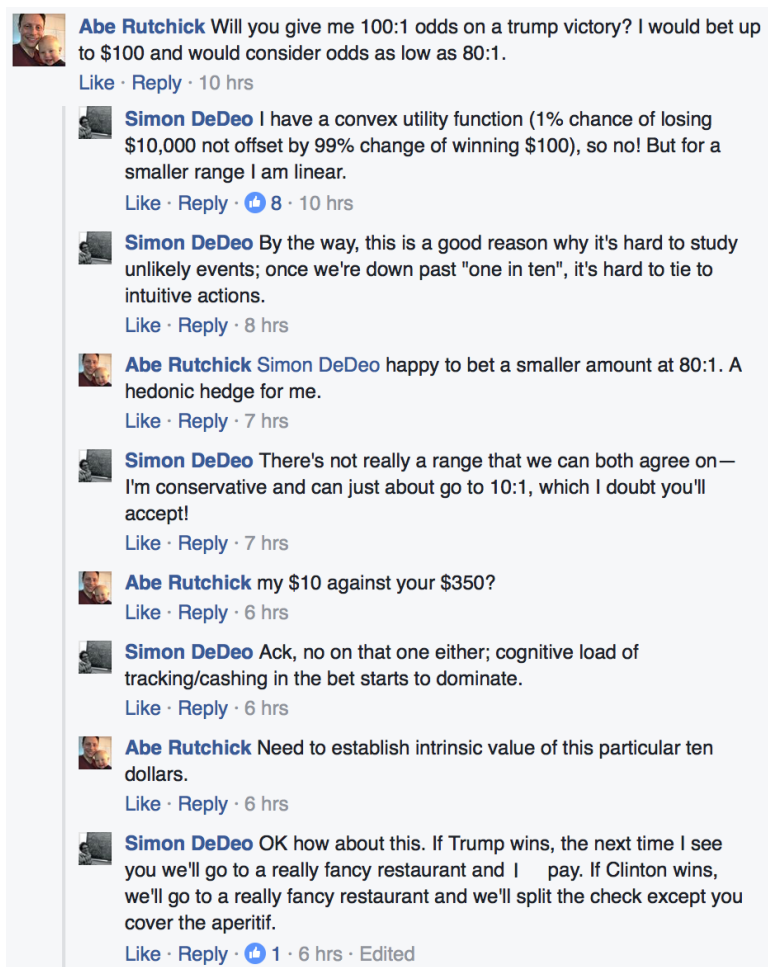


Figure 1: Contrary to what you might learn in an introductory economics class, translating degrees of belief into action is harder than it appears. There are all sorts of reasons why I might believe that an option is only 1% likely to occur, but refuse even to take a 100:1, 80:1, or even 35:1 payout, depending on the amount in play. If we're serious about introspection, it gets harder still—the value of buying my friend a meal seems untranslatable into a dollar amount, meaning that I (Simon) can't think of a strict monetary payoff that would be equivalent to the final deal I offer my friend Abe.

true, but you pay me \$90 if it's false, because “on average”, if we play the game many times, you'll be making nine cents a game (to get nine cents/game, look at it this way: in one hundred games, you'll win \$1 about 99 times, and lose \$90 once).

There are plenty of reasons why you might not want to take this bet, however, even if you hold that degree of belief. One example is if I find it more painful to lose a lot of money than to gain just a little, even if I win “on average”. The relationship can be broken in the other direction, too—I might be happy to wager a small amount for a chance to win big (the lottery ticket story—a one in a billion chance to win a million dollars). We think about people who play the lottery as being irrational, but are they? Say you only have a dollar, and you were starving to death in a ridiculously oppressive country where the cheapest thing is a can of beans that costs \$2. Facing immanent starvation, you wouldn't be crazy to bet your dollar in a game with a 1/3 chance of winning \$2, even though “on average” you lose. A dollar in that world is worth nothing, but two dollars saves your life.

The complexity of cashing out numbers as actions was brought home to me in discussions

concerning the 2016 election (see Section 9 for more). I estimated my degree of belief in a Clinton victory at about 99%; my friend Abe Rutchick then offered me a bet. The resulting discussion is shown in Fig. 1. Some readers might consider me cowardly to reject some of Abe’s offers. That’s fine, since I don’t consider gambling a sign of virtue; indeed, it just goes prove my point because I do, introspectively, hold a very high degree of confidence in Clinton’s win, but you can’t induce me to place the right kinds of bets to prove it. Not just risk aversion, but the incommensurability of values comes into play: I am (introspectively) willing to pay for an expensive meal with Abe, even though I can’t put a dollar value on it. John Searle puts this very nicely, in his 2001 book *Rationality in Action* [2]:

it seems to be a strict consequence of the axioms [of decision theory, the translating of probabilities into actions] that if I value my life and I value twenty-five cents ... there must be some odds at which I would bet my life against a quarter. I thought about it, and I concluded there are no odds at which I would bet my life against a quarter, and if there were, I would not bet my child’s life against a quarter. ... I argued about this with several famous decision theorists, starting with Jimmy Savage in Ann Arbor and including Isaac Levi in New York, and usually, after about half an hour of discussion, they came to the conclusion, “You’re just plain irrational.” Well, I am not so sure.

One of the many things that makes betting your life on a quarter crazy is that if you lose, there are no more bets for you. This terminates the process unexpectedly and makes averages harder to compute. The more general (and less fatal) case of losing everything—and thus being unable to get back in the game to make your losses back—appears in less extreme circumstances where it further frustrates the thinker and requires careful distinctions between “averages over many players at one time” and “averages over one player over many times”; see my colleague Ole Peters on the classic “St. Petersburg” game [3]. When it comes to unique events, like elections, or even historical facts, where repeated bets can *not* be made, the contortions become even worse.

If actions in the real world can’t measure these hidden degrees of belief, are they really real? For many people in the hard sciences it’s the only story we have, because if we want to take states of mind seriously (it is felt) we have to describe them mathematically. We go, then, to the laboratory with the quantification we have, not the quantification we want. A Bayesian story might be good enough, or at least a starting point for something better.

It’s worth, however, looking at what happens when people try to go further. John Maynard Keynes, the twentieth-century British economist, was famous for putting beliefs—the cognitive states of participants in the marketplace—at the heart of both his economic theories and his policy recommendations. Businesses could fall into a self-fulfilling pessimism: refusing to spend, firing workers, and thereby, in the aggregate, depriving themselves of the very markets they needed for their goods. Keynes was no stranger to the measurement of mind; in his early career, before becoming an economist, he had studied probability from a philosophical point of view as a model of belief formation; his first book was *A Treatise on Probability*. The stories Keynes told there is far richer than the Bayesian account presented here; in particular, Keynes presented a coherent mathematical system in which the attachment of a number to a probability, or degree of belief, was neither necessary nor even always possible.

Keynes’ example, appropriately enough for someone writing in the United Kingdom, concerned the question of whether or not it will rain. A Bayesian can always attach a degree of belief to the proposition that it will rain. If we are very uncertain, then it might be close to 50%, but we can not, coherently, refuse to do so. Any story we tell about the ways in which it could rain will, when phrased in the machinery allow us to extract a number, between zero and unity, that corresponds to



the degree of belief we have. Say we think there’s a 5% chance that there’s a drought, and if there’s a drought, then there’s only 1% chance it will rain; and then that there’s a 95% chance of normal weather, for which the chance of rain is 60%; the overall chance of rain is just over 57%. Keynes disagrees with the idea that this can always be run, and in a widely-quoted passage he writes:

Is our expectation of rain, when we start out for a walk, always more likely than not, or less likely than not, or as likely as not? I am prepared to argue that on some occasions none of these alternatives hold, and that it will be an arbitrary matter to decide for or against the umbrella. If the barometer is high, but the clouds are black, it is not always rational that one should prevail over the other in our minds, or even that we should balance them, though it will be rational to allow caprice to determine us and to waste no time on the debate. [4]

Though it might pain the Bayesian to hear me say it, there’s something right about this. Keynes, like Searle, thought deeply, and if you read carefully here you see that he objects not to the ascription of degrees of belief, but rather to the rationality of doing so. Just as John Searle considers some tradeoffs to be irrational and sense-defying, so does Keynes judge some practices of belief-formation. Hidden in Keynes’ objection to the Bayesian account, in other words, is I think a rich *ethics* of thinking. In his youth, as a Cambridge Apostle, Keynes had come to admire the philosophy of G.E. Moore, whose *Principia Ethica* argued that the highest good was an appropriate form of contemplation. No shoving numbers onto sentences for Moore or Keynes! (But no mystification either—Keynes presents an alternate theory where the objects that describe our degrees of belief are more complex yet.)

## 5 Neutrinos, Cable News, and Aumann’s Agreement Theorem

In March 2011, the OPERA collaboration, based in Gran Sasso, Italy, reported a highly unusual discovery: that neutrinos emitted across the border in Switzerland were traveling faster than the speed of light. The effect of this piece of data was electric: the news rocketed around the Internet and then, very soon after, the science sections of the major newspapers in Europe, the United States, and beyond. And yet, within physics departments, it barely passed without a blip—the fuss was even, I venture to say, a bit embarrassing. Why did the same piece of information induce such different subsequent belief?

The answer goes back to the *priors*— $P(T)$ . The Gran Sasso result was in direct violation of the theory of special relativity. Most physicists have very strong beliefs in special relativity; they form these both on the basis of experiments, and the general coherence of the theory itself—what William Whewell first called “consilience” [5].<sup>7</sup> Not only does special relativity predict the outcomes of particle scattering experiments in a lab, but it plays a central role in the explanation of phenomena in a vast number of extremely disparate fields of physics. (There it goes by the more sophisticated phrase “invariance under  $SO(3,1)$ ”.) The response of nearly every physicist on the planet was the same: there must be something wrong with the OPERA experiment. (And, indeed, there was—a cable that was not plugged in tightly enough, causing a very slight delay in timing signals that mimic’d the appearance of superluminal velocity.)

---

<sup>7</sup>Suggestion for a Ph.D. Thesis: a Bayesian theory of consilience. UPDATE this has been done: <https://sci-hub.se/10.1016/j.tics.2020.09.013>.

By contrast, the science journalism community did not have as strong priors on special relativity. Many (though far from all!) science journalists come out of journalism school, rather than science departments. So the strength of their belief in special relativity is based upon the extent to which they trust, without other evidence, the testimony of a few physicists. When confronted with an enthusiastic PR officer (say) from Gran Sasso, and perhaps a few J-School theories about pulling back the lid on powerful secret interests, the balance of certainty shifted enough to write the articles. (There’s an implicit utility function here—the penalty for being wrong, versus missing the scoop—that turns belief into action, but this is the domain of game theory, and that’s another article altogether).

Note that—as in the Artemy and ESP case—physicists didn’t need to think the OPERA collaboration was particularly dysfunctional. We only needed to believe more strongly in the truth of special relativity. Some physicists (and other observers, of which I count myself one) did not discount the result as much as others—not because we thought the OPERA collaboration was particularly good, nor because we doubted special relativity, but because we wondered if there was some feature of special relativity that we hadn’t yet understood, and that would allow us to accommodate the (still revolutionary) Gran Sasso result without losing the coherence of the rest of the theory.

A third example, now that you’ve gotten the hang of it. Say you turn on the TV and hear (for the first time, cast your mind back) the newscaster say “Obama is a Muslim”. What do you believe? In cartoon form—and Bayesian accounts of human cognition always have a cartoony feel—you have a prior about the reliability of the television station, and the possibility that the president of the United States has been maintaining a fictional religious identity in order to conceal a secret agenda. Depending on these priors, rational thought—at least, rational as defined by the criteria on the first paragraph of the first page—will lead two people to very different beliefs, depending on the state of their priors

Note a somewhat sinister aspect here, not present in the previous examples. If it is the case that Obama is a Muslim, and therefore that the president of the United States has a secret religious identity, and nobody else is reporting this, this implies a conspiracy of such vastness that most elite sources of information are probably also implicated. A Bayesian would adjust their beliefs about the reliability of other news organizations downwards, in response; it may not make you believe the news station more, but it will erode your faith in journalism in general.

All three examples go to show that optimal Bayesian reasoning need not lead people to agreement. This is contrary to how we often try to settle arguments, where we may want to say “we disagree; let’s all sit down and talk this out”—in the spirit of Leibnitz, “*calcelemus*”.<sup>8</sup> As our examples here show, such disagreements may *never* be resolved, if the priors are too different. If the physicist can convince the J-School student not to report the Gran Sasso result (just yet), it is in part because they have gone back to earlier assumptions where the priors are sufficiently strongly shared. (As a contrary example, imagine someone whose priors are—for some reason, say having to do with your sex, race, or religion—that you are fundamentally unreliable. No amount of evidence you provide will ever allow you to overcome even the slightest bias that person holds.)

This is a sad story. Peircean and Popperian accounts of a mystical future agreement among all rational agents *must* supplement Bayes with something—even if only that we, at heart, share some truly fundamental priors.<sup>9</sup>

---

<sup>8</sup>The fact that this never really works anyway means, of course, that the application of Bayesian models to human cognition has at least one successful prediction under its belt.

<sup>9</sup>The distinction between a prior and a posterior is a fluid one. Most things I might describe as a prior—such as my belief in the OPERA collaboration’s reliability—are secretly beliefs influenced by other forms of evidence, which

Oddly enough, if you do make that jump, then *the only* thing we need to do is agree on our priors. An amazing theorem, due to the economist Robert Aumann, says that (1) if two agents are rational (follow the Bayesian rules of this paper), (2) if they have the same priors, and (3) if they have common knowledge about each other’s beliefs, then they “cannot agree to disagree” (*i.e.*, they must agree). Put poetically, say I encounter a child from the slums of New Delhi. She and I have had completely different life experiences; we have seen different things, learned different things, and now (because of those experiences) believe different things. If she and I are both rational agents, however, and share the same priors, and *if we discuss long enough*—“hmm, I believe this, but you don’t, but now I see that you must believe the opposite because of this, but I don’t believe your evidence for that because of this, but I do adjust my belief here what do you think about that?”—we will, according to Aumann’s theorem, come to share the same views.

This is true even if we think the other person is an unreliable witness! It works, in other words, even if I hold beliefs about her mis-remembering or mis-reporting evidence. As long as I somehow know what her beliefs are, and she knows mine, and I know she knows mine, and so on, and we agree on our priors, we will come to agree.<sup>10</sup>

I’m not even kidding [6]. In the words of Aumann’s abstract, which spends most of its time defining the discussion bit (common knowledge),

Two people, 1 and 2, are said to have common knowledge of an event  $E$  if both know it, 1 knows that 2 knows it, 2 knows that 1 knows it, 1 knows that 2 knows that 1 knows it, and so on. THEOREM. If two people have the same priors, and their posteriors for an event  $A$  are common knowledge, then these posteriors are equal.

where “posterior” is just fancy for “the beliefs you have after updating your priors given your (idiosyncratic) experiences”— $P(T|D)$ . Better yet, for optimists of the human condition, the computer scientist Scott Aaronson showed that it may not even take that long [7].

The real problem—trauma—is the need to have priors that agree. What do you believe before you start getting data—before, in other words, your experiences begin? An optimist might say that evolution would give us all similar priors; a pessimist would point out that it would make the subsequent intergalactic war, with aliens that evolved under very different conditions, that much more violent.

The jury is out. Because of the optimality results (page one, paragraph one) Bayesian reasoning is often used in artificial intelligence and machine learning algorithms. Indeed, if we had enough computer power, that’s exactly what we’d do; when we don’t use Bayes, it’s because it can be expensive to compute and the machine takes shortcuts.

But it goes beyond the world of machines, because it also seems to describe human behavior—or, at least, to provide interesting thought experiments that illuminate features of human behavior in a

---

themselves depend on other priors, which, on further examination turn out to be themselves derived from priors, *ad infinitum*. Practically speaking, we roll with the priors we have, cracking them open when we find ourselves in disagreements with other rational beings. The origin of the priors themselves is mysterious. Examine Eq. 2 again: one never, in the Bayesian framework, intuitively believes—one only updates them, going from  $P(T)$  to  $P(T|D)$  to  $P(T|D, D')$  to... Reverse that order. Can it be “evidence all the way down”? The standard answer is no: that priors at some point are fixed by extra-logical forces such as evolution.

<sup>10</sup>Important note: we do have to be honest about our beliefs! We may think the other a scoundrel, but if the theorem is to work, we must respect them enough to tell them so. I am glossing over a hole in my use of Aumann’s theorem, which is how we gain knowledge of the other person’s beliefs. Aumann’s theorem says that once the beliefs become magically known, they will eventually (after discussion) come to be shared. If we are good Bayesians, we’d have to build a theory of our partner’s beliefs based on the words they emit and gestures they make when they try to tell us.

new way. It can explain why the more the person at the café tells you about their ESP experiences, the less and less you believe them; it can explain why rational people can disagree, when given the same data, and even why, as you give them more and more shared experiences, they might draw further and further apart.

It does this explaining by reference to mathematically optimal reasoning, which has a pleasing feel. Simple models seem to show that, even when we're behaving in ways that seem irrational, we may not be so irrational after all. Bayesian accounts of reasoning also suggest that disagreements might be resolved by argument and discussion of an appropriate form, rather than appeals to emotion, or violence. They even tell you some things about what that appropriate form is. It's not a Pollyanna theory, where things can never go wrong, but it begins to specify the conditions under which the course of human history might go right.

## 6 Specifying Priors and the Zen Koan of Marvin Minsky

The role of priors in Bayesian reasoning is not uncontroversial: by adjusting her priors, a scientist can reverse the implications of an experiment. But this is not as bad as it might seem. In general, there's a sense that as long as priors do not attribute zero probability to the “true” theory, a sufficient number of experiments, of sufficiently wide range, will—eventually—overwhelm priors biased against reality. The scientist who conveniently adjusts her priors will be wrong in the long-run.

As the ESP example shows, though, there are limit cases. It tends to feel, however, as if these limit cases depend for their effect on artificially restricting the space of theories. We do, for example, have many different methods of assessing Artemy's sanity; we can send additional observers; we can conduct new experiments; and a sufficient number of these should, in the final analysis, be sufficient to overwhelm even the strongest biases. If Artemy had reported the results of a test of Bell's Inequality, before I had Quantum Mechanics, I'd be dragging my feet, and maybe even a bit rude, too—but not unconvincible.

Weasel words, however, abound, when we try to answer these questions: “should”, “final analysis”, “sufficient number”. I am aware of no theorem that demonstrates the asymptotic independence of belief from “reasonable” priors. Nor, indeed, do I have a definition of what it means for priors to be reasonable.

Modulo this rather scientific faith, our lack of understanding places us, if provisionally, in a strangely Cartesian place of absolute doubt, where the dependence of our beliefs on our priors means that vast systems of contradictory beliefs co-exist.<sup>11</sup> Indeed, we are in a place even worse than where Descartes found himself. If we were only uncertain about the world, we would have a well-defined system of beliefs, most of which are around 1/2; but priors are sufficient not only to turn certainty into doubt, but doubt into certainty. (Why not factor this in? If different priors give different beliefs, why not put priors on the priors, and average? In the literature, this is known as a “hyperprior”—and only, of course, punts the problem one step back.)

None of this invalidates the Bayesian project; indeed, one might phrase the result as the *discovery* of priors. The fact that we have been doing statistics without talking, or even knowing, about them doesn't mean they didn't exist. They were there all along, lurking in implicit and incoherent form for as long as we had been reasoning: a state of innocence, rather than grace.

An “AI Koan”—a parable, half-joking—from the early days of artificial intelligence research

---

<sup>11</sup>David Deutsch, in *Fabric of Reality*, provides a beautiful example of the scientific faith—including an example of how, by sheer force of intellect, he battles his way out of a “brain in the vat” virtual reality; see Chapter 5.

summarizes this position. Attributed to the computer scientist Danny Hillis, it tells the story of two giants of the field, Marvin Minsky, and his student Gerald Sussman, from the early 1960s:

In the days when Sussman was a novice, Minsky once came to him as he sat hacking at the PDP-6.

“What are you doing?”, asked Minsky.

“I am training a randomly wired neural net to play Tic-Tac-Toe” Sussman replied.

“Why is the net wired randomly?”, asked Minsky.

“I do not want it to have any preconceptions of how to play”, Sussman said.

Minsky then shut his eyes.

“Why do you close your eyes?”, Sussman asked his teacher.

“So that the room will be empty.”

At that moment, Sussman was enlightened.

“A Selection of AI Koans”, *New Hacker’s Dictionary*; Ref. [8]

## 7 Further Reading

If you're interested in the quantification of belief, and questions like how beliefs change over time, and how beliefs relate to, and influence each other, you will want to learn some information theory. See the companion article, *Information Theory for Intelligent People*, <http://santafe.edu/~simon/it.pdf>.

The physicist E.T. Jaynes died in 1998; his magnum opus, *Probability: the Logic of Science* was reconstructed from his typescript notes (that appear here as Ref. [1]). Rather mathematical, it provides the modern foundation of the use of Bayesian reasoning—and it also provides the ESP example for this article (in the chapter “Queer Uses for Probability Theory”). Rumors suggest that the polemical and harsh nature of the book in parts was due less to Jaynes himself than the passions of his disciples—a Christ and St. Paul scenario.

The physicist David MacKay, while not a Jaynesian, wrote perhaps the best sequel to Jaynes' book, again, highly mathematical, and appropriate for the modern AI era, called *Information Theory, Inference, and Learning Algorithms*; it's available free at <http://www.inference.phy.cam.ac.uk/itprnn/book.pdf>, but also in a lovely hardback. David was a much-loved man of reason, knighted by the British Queen, made a Fellow of the Royal Society, and, when not writing clearly for us, advising the British government on sustainable energy, and raising a family—all by the age of 48, when he died, far too young, of stomach cancer.

The physicist David Deutsch, spot the pattern, has written a great deal on the role—and limits—of Bayesian inference, in popular writings that are just about as bizarre (and fun) as Charles Sanders Peirce (physicist). His books *The Fabric of Reality* and *The Beginning of Infinity* pick up where Karl Popper (not a physicist) left off, and build an entire metaphysics around the use of reliable reason and hypothesis construction.

## 8 Technical Notes on Working with Probabilities

Once you get the hang of it, Bayesian probability is addictive. The notation for probabilities varies from person to person and field to field, but fortunately there are some general rules. By convention, when we write  $P(x, y)$ , the probability that  $x$  and  $y$  are both true, the event  $x$  is drawn from a set of possibilities we label  $X$ , and  $y$  from a set of possibilities we label  $Y$ ;  $X$  and  $Y$ , once we attribute probabilities to their sentences, are often called “random variables”.

Each set must *exhaust* all the possible values for that variable, and be *mutually exclusive*: for example,  $Y$  could be a trio of sentences, {“the student chosen at random is male”, “the student chosen at random is female”, “the student chosen at random is neither male nor female”}. “Exhaustive” can be a subtle concept: if, for example, we are working from an archive with missing entries, then  $Y$  might require a fourth sentence: “the student’s gender is unknown”.

Many errors in working with probabilities have their source in ambiguities; most commonly, ambiguities involving the sentences that make up your sets. Consider the following scenario: Cody Zeller<sup>12</sup> is near a Penn State player in the final quarter; the Penn State player trips and falls. You are curious about how the Penn State player being on the ground is related to the possibility that Zeller fouled him.

You might consider having two random variables to describe the situation. One, call it  $X$ , might have two sentences, {“Zeller fouled him”, “Zeller didn’t foul him”}; the other, call it  $Y$ , might be {“Penn State guy hasn’t tripped”, “Penn State guy has tripped”}. Because each of these sets separately exhausts the possibilities, they’re good.<sup>13</sup>

Or you might consider one random variable,  $Z$  that has four sentences, combining all the different possibilities, such as “Zeller fouled the guy and the Penn State guy hasn’t tripped”. But if you tried to reason from a variable  $Z$  that had two only two sentences, {“Zeller fouled and the guy tripped”, “Zeller didn’t foul and the guy tripped”}, you would later find yourself in confusion—it is entirely possible that Zeller did foul the guy, but he caught his balance and didn’t trip, and by excluding this from your set, you will get incorrect answers when you grind through the mathematics.

Or, if you had two sentences, {“Zeller fouled”, “Zeller fouled and the guy fell”}—in this case, the sentences are not mutually exclusive (both could be true at the same time). Often a sign you’ve messed up here is that your probabilities for all the items in a random variable don’t sum to unity.

One of the trickiest things is how the use of probabilities violates how your math teacher taught you to work with functions. In high school, you might encounter a function  $f(x)$ , which could be, say, the quadratic  $x^2 + x + 1$ . You were able to change the argument, so you could write  $f(c)$ , and everyone knew that you meant  $c^2 + c + 1$ . The “ $f$ ” in  $f(x)$  or  $f(c)$  names that particular functional form.

However, when we write probabilities, we use the same “name”,  $P$ , for everything. We rely on the variable name to tell us which one;  $P(x)$  means the probability of event  $x$ ;  $P(y)$  the probability of event  $y$ . But what if we want to use a variable for events? For example, say we want to take the joint probability  $P(X, Y)$ , and sum over all the possible sentences in  $X$  to get the probability of  $Y$

---

<sup>12</sup>A one-time basketball player for Indiana University, a wonderful institution that, without its knowledge and thanks to the graciousness of the American people and the citizens of Indiana, paid for me to write this.

<sup>13</sup>Here you know they do because of something pretentiously called the “law of the excluded middle”—which says that “either  $x$  is true, or not- $x$  is true.” Once something obvious is called a law, you have the benefit of being able to ask what happens if the law isn’t true. Some (to this author tendentious) interpretations of quantum mechanics make use of novel logical systems that violate the law of excluded middle. In a different vein, the logicians Graham Priest and Richard Routley proposed to reject this law in the classical case to construct a new “dialethic” logic (see, e.g., Priest’s *Doubt Truth to Be a Liar*). The extensions of such a system to encompass doubt and uncertainty remain to be done, but would require revision of the basics of probability theory itself.

alone; we might write

$$P(Y) = \sum_{i \in X} P(i, Y), \tag{10}$$

where  $i \in X$  is shorthand for “all of the different sentences in  $X$ ”, and  $i$  is a variable. In the formula  $P(i, Y)$ , it can be easy to forget what the first argument means; it’s often good to cue in a reader by writing  $P(X = i, Y)$ . This gets particularly important when the sentences are naturally written as numbers (*e.g.*, “the price of Apple is \$16/share”)—if you have a joint probability describing two stocks, you can write  $P(12, 19)$ , but it might be better to be careful and write  $P(\text{Apple} = 12, \text{IBM} = 19)$ . We’ll see this in the next section, where we attack a harder problem about waiting for the bus.



## 9 Nate Silver, Sam Wang, and who will win the election?

*Update:* please see “Wrong for the Wrong Reasons”, <http://santafe.edu/~simon/wrong.pdf> for a postmortem of 2016 election polling and statistics.

In 2004, the Princeton neuroscientist Sam Wang began applying Bayesian tools to the prediction of U.S. presidential elections. Polling companies—the people who called up voters, or did internet surveys, or worked by mail or knocking on doors—had begun putting their raw data online. If you wrote down models for how people’s underlying voting habits would be reflected in these polls, then the same models would also predict who would win the elections. Sam built some simple models (or, rather, wrote down some nice stories about  $P(D|T)$ , where  $T$  were the underlying voter preferences, and  $D$  was the polling data), and predicted the outcome of the 2004 presidential election on the nose.

In 2008, Sam was joined by Nate Silver; Nate had learned the power of Bayesian tools as a sports statistician.<sup>14</sup> Together with a number of other hackers and part-timers, they changed the face of election prediction, taking it out of the hands of the pundits—who often inflated uncertainties in order to create a television narrative—and putting it back into the hands of the people. This new kind of psephology, with Bayes at the heart, is more accurate than anything we’ve had before, helping wonks and nerds, and ordinary citizens, focus their attention on what matters and where the most is at stake.

How do Nate and Sam do it? Let’s take some steps towards a Bayesian model of election prediction. In doing so we’ll get a few toy examples of how Bayesian statistics is actually used in practice. Let’s assume (for the moment) that we have good ways to build beliefs about which candidate will win state-by-state. Then a natural thing to do is to combine these to get the degree of belief one should have in a particular combination of outcomes. Say, for example, we want to know the probability that Clinton wins both Pennsylvania (PA) and Ohio (OH), and we have polling data ( $D_{PA}$ , and  $D_{OH}$ ). It’s then natural to write

$$P(PA \& OH|D_{PA}, D_{OH}) = P(PA|D_{PA})P(OH|D_{OH}) \quad (11)$$

where (for notional simplicity) the arguments of the  $P$ s will refer to Clinton (*e.g.*,  $P(PA|D_{PA})$  means “the degree of belief I have in Clinton winning Pennsylvania, given polling data from Pennsylvania  $D_{PA}$ ”). This is the equivalent move to saying that the probability of getting Heads from two coins is just the probability of getting Heads with coin one, times the probability getting Heads with coin two.

There are three things wrong with doing this. The first is that I haven’t told you how to get  $P(PA|D_{PA})$  or  $P(OH|D_{OH})$ . The second is that (as hinted), we’re assuming that, given polling data, states are independent.

The third is that if we try to scale this method up, we need to compute the probability of every single pattern of winning and losing. When we start doing that we realize that we end up with at least  $2^{51}$  different things to calculate, because we have to consider every single pattern (Clinton does/doesn’t win OH, Clinton does/doesn’t win PA, Clinton does/doesn’t win NV, ...) of the electoral college. Every state (and D.C.) contributes two possibilities—either Clinton wins all the electoral votes (EVs) for the state or not; we’ll neglect for simplicity states like Maine where they hand out the EVs separately. Even if we can compute one probability every microsecond (a

---

<sup>14</sup>I’m pretty sure that I overheard Nate explaining Bayes’ theorem while waiting for the 55 bus in Chicago one very cold night in 2008, but that’s a story for Section 10.

standard lag time for code like Python), it will take seventy-one years to compute them all, and 2,000 Terabytes just to store the results.

Begin with this third problem, which turns out to be the simplest to solve. On reflection, there are three ways out.

You could say, well it’s not the case that we have to consider every possibility—there are many states that Clinton is practically guaranteed to win, so we shouldn’t bother computing things like “the probability that Clinton loses New York but wins Ohio”. She won’t lose New York. If there are just ten swing states, we can get by with  $2^{10}$  different combinations—about a thousand—and that’s pretty quick. Or (a second idea): we can simulate an election, going through the 51 electoral votes and tossing biased coins for each state to get a sample of what might have happened—a kind of *SimCity*, or Sid Meier’s *Civilization*, writ small. If we do this lots of times, and keep track of the final votes, we can build up a probability distribution for anything we care about (*e.g.*, and most simply—we can simulate the election 10,000 times and see how many times Clinton wins). This second method is a very simple example of what the stats gurus call “Monte Carlo” (after the famous gambling city in Monaco). Monte Carlo solves a Bayesian reasoning problem by doing what all Bayesians secretly fantasize about doing: make it into a (very controlled example of a) frequentist sampling story. Monte Carlo is pretty good because (roughly speaking) if you run 10,000 simulations and a particular outcome never occurs, your degree of belief in that outcome happening in the real world should be no more than around  $10^{-4}$ .

A third method is a little trick that Sam Wang is particularly proud of [9]; it turns out that if all you care about is the number of electoral votes—and not where they come from—you can get that really quickly. First, write down the polynomial

$$f(x) = \prod_{s=1}^{51} [(1 - P(s|D_s)) + P(s|D_s)x^{EV_s}], \quad (12)$$

where  $s$  is a variable for the states,  $EV_s$  is the number of electoral votes coming from state  $s$ ,  $P(s|D_s)$  is the probability of Clinton winning state  $s$  (given the polling data  $D_s$ ). What’s  $x$ ? Don’t worry about it, just write out Eq. 12 blindly, like a robot with no soul (*i.e.*, focus on the syntax of the formula).

Sam’s trick is to note that the coefficient of  $x^N$  in the function  $f(x)$  is equal to the probability that Clinton wins  $N$  electoral votes. That only requires you to compute a polynomial by multiplying things together 51 times! If you’re a math geek, you love this and can see why it’s true; if not, try writing out the product in full for a simple case with, say, four states, and you can see that all the cross-multiplications you’re doing amount to considering the  $2^4$  different possibilities, except that as you add together terms that have the same power of  $x$ , you’re no longer keeping track of them separately. This is a simple example of something the gurus call a generating function.

Looking at the simplest problem is a nice way to get a sense of the “technology” behind Bayesian computations. You can get an exact answer to a simpler question by hand; you can approximate the full question with Monte Carlo simulations and an expensive computer; or you can be super-clever (sometimes) and answer part of the question exactly and everyone’s impressed.<sup>15</sup>

The second hardest problem is the independence one. Is it really the case that “Clinton winning Ohio” and “Clinton winning Pennsylvania” are independent events? In one sense, of course not. If Clinton is losing Pennsylvania badly, she’s likely to be losing Ohio, too. But that’s not what Eq. 11

---

<sup>15</sup>In graduate school, these were associated for me with different nationalities. The first method seemed British (pragmatic compromise), the second rather American (throw money at it); the third very Soviet Russia (uncompromising hard-slog brainpower).

is saying, because  $P(OH|D_{OH})$  is the probability of Clinton winning Ohio *given* the polling data. If Clinton is losing Pennsylvania badly, then hopefully the reasons that she's losing Pennsylvania badly, and that mean she'll also lose Ohio, will show up in the Ohio polling data. Under this assumption, the Ohio polling data will move up and down with the Pennsylvania polling data. But once I know the Ohio polling data, any further motion in Pennsylvania will be irrelevant to predicting Ohio.

An analogy would be the example of two friends who go to a bar together, and drive home separately. Their drinking is correlated—if Alice starts drinking a lot, then Bob is likely to join her. This means, therefore, that their probabilities of getting into a car accident are not independent: the probability that Alice and Bob *both* crash their cars is higher than the probability that Alice crashes times the probability that Bob crashes. But once you tell me the blood alcohol content (BAC) of Alice, that's all I need to know to predict the chance that she'll crash her car. Formally, it's the case that Alice and Bob are strongly correlated

$$P(A, B) \neq P(A)P(B) \quad (13)$$

but become independent once you know their BACs,

$$P(A, B|BAC_A, BAC_B) = P(A|BAC_A)P(B|BAC_B). \quad (14)$$

The pretentious way to say this is that sometimes variables are conditionally independent. When you build models of many variables, and specify the various patterns of conditional independence, you're well on your way to building what's called a Bayesian network, which is related to what Judea Pearl calls a Causal Network, and is the basis of a beautiful and complex and (in this author's opinion) flawed theory of measuring causality from empirical data, but that's another story.

If Pennsylvania and Ohio going to the polls are like Alice and Bob drinking after work, then Sam's independence assumption does very well. But you can imagine ways in which it goes wrong. The first thing you might say is that perhaps polling data from Pennsylvania really does tell you something about Ohio that Ohio data doesn't tell you. That turns out not to be so interesting, because we can recover independence now just by folding the Pennsylvania data to our Ohio prediction: formally, figuring out how to write down  $P(OH|D_{OH}, D_{PA})$ .

To *really* break independence, you need some causal relationship that can't be captured by polling. Imagine, for example, that bad news for Clinton in Pennsylvania depresses Clinton-voter turnout in Ohio (perhaps easier if you think about, say, Colorado, where timezones mean the Coloradans can watch the exit polls from Pennsylvania and decide whether or not to vote).

A more abstract possibility is that of *correlated errors*. In order to compute  $P(PA|D_{PA})$  we have to turn polls of individuals into predictions about state-wide events. To do this, we have some theories—for example, we might decide that women are more likely to vote, and so  $P(PA)$  is higher than you'd expect because women favor Clinton. What if those theories are wrong?

To continue our example, say that we're wrong about women having high turnout. In that case, we not only over-estimate the probability of  $P(PA|D_{PA})$ , but we're wrong about every state. To take this into consideration, we have no choice but to model everything together, including different theories about turnout, and writing something like,

$$\sum_{T_w=1,2,\dots}^N P(PA, OH|D_{PA}, D_{OH}, T_w)P(T_w) \quad (15)$$

where  $T_w$  is a (discrete) variable representing  $N$  different theories about turnout. Now the outcomes in Pennsylvania and Ohio are correlated again, because they both depend on the value of  $T_w$ . To

go back to Alice and Bob—allow me this slightly weird example, it’s late at night before the 2016 election—imagine that they’re related to each other. Because they’re related, they might both have genetic sensitivities to alcohol. If they do, even low BAC levels lead to significant impairment. Now all of a sudden, their futures become correlated again, even once we know their BACs; if Bob has low BAC but crashes his car, then we say “oh no—now that I know Bob just crashed with low BAC, I bet Alice has the same bad gene.”<sup>16</sup> The big debates among the stats gurus at the end of the 2016 election center around this possibility: Nate says you have to consider it, Sam says it doesn’t matter.

Talking about turnout models leads us, finally, to the question of how to turn polling data into “state win probability data”—the  $P(PA|D_{PA})$  that form the basis of the final prediction. This is, in a sense, where all the magic is. In general, Nate has very complex models, while Sam has very simple ones. Indeed, in Sam’s 2012 paper, he reports a rather heuristic method that (wait for it) involves taking the medians of all the polling data for the state, and pretending that that median is normally distributed with a variance  $\sigma$  given by the median of all the pollster-reported errors, divided by  $\sqrt{N}$ . There’s some justification for all of these steps, which in the end give you a  $P(PA|D_{PA})$ , but it’s not a simple Bayesian story any more—at best, it’s a sort of heuristic approximation to the real story you *should* be telling. If you’re a statistical type reading Sam’s report of what he does you sort of nod along, half-filling in the gaps as to why each of these steps is a good approximation, but it’s not the world’s most rigorous demonstration.<sup>17</sup>

Nate has much more complicated models, that involve building careful stories about voter turnout, pollster biases, and so forth. The short version of what this all means is that Sam tends to be much more certain than Nate in his predictions. For Sam, once you write down the state-by-

---

<sup>16</sup>A common genetic defect like this does indeed exist. It prevents the creation of alcohol dehydrogenase, the enzyme that processes alcohol. Without alcohol dehydrogenase, drinking even small amounts of alcohol leads to the buildup of toxic byproducts that cause nausea, mental confusion and blurred vision—instead of a pleasant “buzz”. But is this defective gene “bad”? Actually, alcohol is a terrible thing for your reproductive success, because drunk people make bad decisions. If you could evolve a desire never to drink it, your descendants would do better. Evolution works on slow timescales, but it has had time to act here as well. The first cultures to cultivate a grain easily fermented into alcohol were in Asia, where it first became possible to make rice wine. Today, defective alcohol dehydrogenase genes are preferentially found in East Asian populations, including the Han Chinese, which also have lower rates of alcohol abuse. College students with ancestry from other parts of the world refer to this condition as “Asian flush”, perhaps unaware that the joke’s on them. The relevant timescales (10,000 or so years since rice cultivation) are just about the shortest that we know evolution can work on in human populations—which is why later wheat cultivation hasn’t had time to flush it out of Europe (yet).

<sup>17</sup>Savvy appreciators of Sam’s analysis will see how much I’ve left out. To begin with, Sam now *does* have a model to account for correlations in polls: what he calls the meta-margin (MM), which describes the possibility of a uniform bias across all polls in the direction of one candidate or the other. Hypotheses about what this meta-margin is going to do lead to correlated changes in  $P(s|D_s, MM)$ , which he can then turn into EV predictions using the generating function trick,  $f(x)$ .

Sam’s tried a few theories about the MM over the years, but in the most recent election, 2016, he used past data to formulate a prior that assumed a symmetric distribution of the meta-margin about zero—*i.e.*, while the polls will, in the end, be biased in one direction or another, he remains agnostic about which direction they’ll be biased in. To build his final predictions, he combines a Monte Carlo method with his generating function trick: he samples from the prior distribution over MM (the simulation step), gets a bias, feeds that into  $P(s|D_s, MM)$ , and gets an EV prediction. He repeats this a few thousand times to produce a distribution over different EV totals, and uses that to get probabilities on which candidate will win. The MM prior reduces the certainty of the final result, but a quick analysis suggests that it can never actually flip the prediction of who is most likely to win the election. At other points in his analysis, Sam has used the MM not to model polling bias, but rather opinion drift in the underlying preferences of likely voters. It’s a bit of a hack, but assuming a prior distribution over MM can be thought of not as a theory about polling errors, but as a story about actual voters randomly flipping their opinions back and forth in a noisy, but correlated, fashion from moment. I thank Joel Erickson for delving into Sam’s 2016 methodology with me.

state probabilities, it's pretty much a foregone conclusion, because even slight leans in each state accumulate in favor of a candidate. All you need are enough polls to make those leans apparent enough, and near the end of an election, that tends to be the case.

For Nate, it's very different. Correlations between states mean that a single hidden variable (say, young female turnout, variables like  $T_w$ ) can wash across multiple states at once, erasing mild leans that favor one candidate. If you're uncertain about the state of those hidden variables, you're much less certain about the state of the election as a whole.

Who does better? Both Sam and Nate have logged great successes, beginning in 2004 when Sam's raw code predicted not just the election, but the exact EV split using his generating function  $f(x)$ . It's also worth noting that both Sam and Nate have had forehead-smacking moments when they step in to override the results of their more naive models. Sam, for example, erred in 2004 when he "added" likelihood to Kerry to handle what he believed would be late-breaking voters for the Democratic candidate John Kerry—some wishful thinking natural to a committed Democrat.

Meanwhile, Nate stumbled in 2016 when, in a more complex series of events, he failed to predict the nomination of Donald Trump. Prediction markets—*i.e.*, sites that allow punters to place bets on election outcomes—had favored Trump for many months before he won the nomination. But while the clever aggregation of lots of individual beliefs of many thousands of election junkies did very well, the pundits mostly "could not believe it", even as the evidence, at least as perceived by the wisdom of the crowd, piled up. Nate wrote a nice *mea culpa* on how he got caught up in the pundit's game, letting his prejudices and gut feels take over from the mathematics.

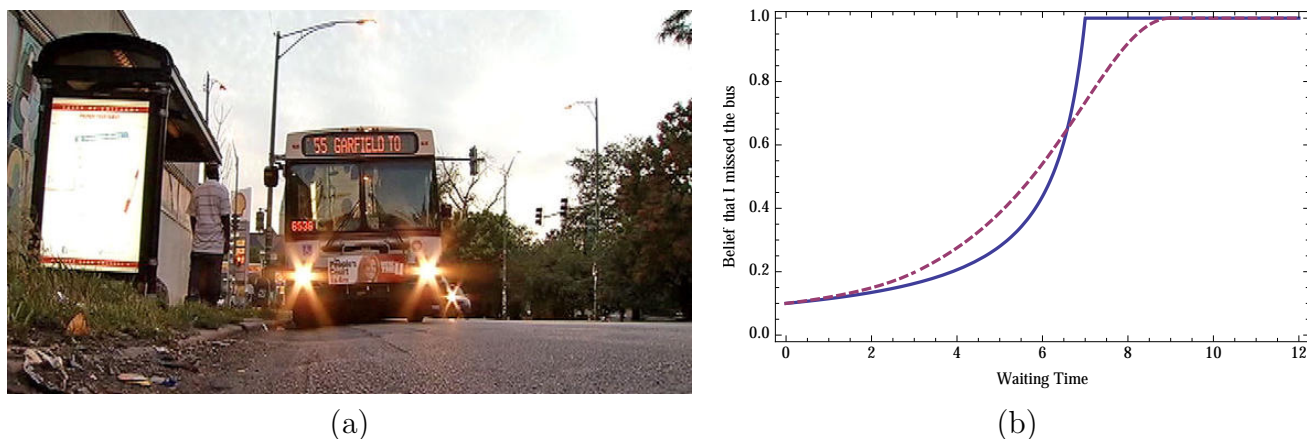


Figure 2: **Left panel:** the Number 55, feared and loved by generations of students at the University of Chicago, as their only connection to the rest of the city. When will it come? Image from NBC Chicago, 26 August 2013. **Right panel:** Did I miss the bus? What a Bayesian angel believes after waiting for  $t$  minutes, given two different theories she might have about bus schedules.

## 10 Waiting for the Number 55 bus

In this (advanced) appendix, we’ll use a little mathematics to show how a Bayesian angel waits for the bus.

For concreteness, we’ll take a scholarly example. The University of Chicago, in Hyde Park on the South side of Chicago, is singularly ill-connected to the rest of the city. Students are forced to rely on “the 55”, a city bus that picks up along 55th Street and connects a few miles away to the subway system. Returning home late at night, students wait outside the station for the ride that takes them home. But when will it arrive? When temperatures drop, as they do in Chicago, below zero degrees Fahrenheit, the question becomes an urgent matter of psychology.

Intuitively, you don’t expect the bus to be there right away. Though it might come quickly, you assume you’ll have to wait a bit before it shows up. However, as the minutes tick by, a mounting despair takes over. Is the traffic bad? Did the bus break down? Has the schedule gone irregular? Has the last bus of the night come and gone?

From a Bayesian perspective, a student waiting for the bus is not just waiting for the bus: she is also, simply by waiting there, gathering information for and against different hypotheses about what is happening elsewhere in the city. Each hypothesis, understood separately, predicts a different arrival time. As the student waits, and some of these times pass without a bus, hypotheses are killed off. The information the student gets simply from waiting means that her beliefs about when the bus will arrive—how much longer she has to wait—are in flux.

Let’s see how this works in detail. In particular, let’s plot out the changing beliefs that the student has as she waits longer and longer. To work through this problem in a Bayesian fashion, we’ll do two things. First, we’ll describe the space of *theories* the student can have. Each theory will imply a certain probability that the bus will have already arrived while the student was waiting. As the student continues to wait for the bus, her belief in some theories will rise, and her belief in others will fall.

Some theories—such as the theory that buses have stopped running for the night—will be simple to describe. Others will, as we’ll see, be a little more complicated, and have some hidden knobs and dials or what are known as “adjustable parameters”. Once we’ve defined our theories, we’ll describe the predictions the theories make.

For simplicity, we'll have two theories in our list  $T$ : theory  $a$ , that the buses have stopped running for the night; and theory  $b$ , that the buses run on a regular schedule, spaced exactly seven minutes apart.

Notice that  $b$  is not quite as simple a theory as theory  $a$ ; the buses may be running on a regular schedule, but the student doesn't know whether one just arrived right before she did. So for the student's point of view, theory  $b$  has an adjustable parameter, call it  $\phi$ , the number of minutes before she arrived that the previous bus left. By stipulation,  $\phi$  is between zero and seven. If  $\phi$  is zero, the student arrived just after the bus left (argh!).

As for the data, or list of possible observations  $D$ , we'll describe it by a single number,  $t$ , the number of minutes the student has been waiting; at any point in time, we can ask "what's the probability that you're still waiting after time  $t$ , given theory  $t \in T$ ". (You could imagine adding in additional information, such as the observation of buses going the opposite direction or the presence of other people waiting at the stop, that might bear on the truth of different theories, but we'll keep it simple for now.)

Predictions with theory  $a$  are simple. If the buses have stopped running for the night, the student will necessarily still be waiting, so  $P(t|a)$  is equal to unity (unless  $t$  gets really large, let's say, and morning rolls around and the buses start again—but forget this).

The probability that the student is still waiting if theory  $b$  is true is more complicated, since  $b$  has an adjustable parameter,  $\phi$  ("how recently the last bus came"). We know that  $P(t|b, \phi)$  is equal to unity if  $\phi + t$  is less than seven, and zero otherwise; informally, if it's been less than seven minutes since the last bus came, she's definitely still waiting; otherwise, she's definitely not.

What should we believe about  $\phi$ ? If we're optimists, then  $\phi$  is close to seven; if we're pessimists ("I'm so cold and I'm SO SURE we just missed it I hate my life"), we think it's close to zero. Let's be neutral on the question, and assume our prior on  $\phi$  is uniform:  $P(\phi)$  is the same for all values of  $\phi$ . This is a bit tricky because in general  $\phi$  is a continuous quantity, not a list of sentences; we could restrict  $\phi$  to be an integer, and do sums, but for fun, let's use some calculus to do the limit where  $\phi$  can take on any real value between zero and seven and sums over values are replaced by integrals. Then we have

$$P(t|b) = \int_0^7 P(t|b, \phi) P(\phi) d\phi. \quad (16)$$

If  $\phi$  is in units of minutes, then  $P(\phi)$  will be a constant,  $c$ , which turns out to be equal to  $1/7$ , as can be seen from the normalization condition

$$\int_0^7 P(\phi) d\phi = \int_0^7 c d\phi = 1. \quad (17)$$

Because  $P(t|b, \phi)$  is equal to unity only when  $t + \phi$  is less than seven, we have that  $P(t|b)$  is equal to zero when  $t$  is greater than seven, and

$$P(t|b) = \frac{1}{7}(7 - t) \quad (18)$$

otherwise. As  $t$  gets larger, the range of  $\phi$  for which the integrand of Eq. 16 is non-zero gets smaller and smaller.

We specified a prior on  $\phi$  for the parameter of our theory  $b$ ; now it remains to us to specify priors over  $a$  and  $b$  as a whole. If we feed  $P(a)$  equal to 0.1 (a small chance that the buses have stopped running) and  $P(b)$  equal to  $1 - P(a)$  (*i.e.*, these are the only two possible theories that could be true), we can now describe anything we desire about the student. For example, we can

use Bayes' rule to flip  $P(t|a)$  around into  $P(a|t)$ , the belief the student has that, given she's been waiting  $t$  minutes, the bus is never going to come:

$$P(a|t) = \frac{P(t|a)P(a)}{P(t|a)P(a) + P(t|b)P(b)} \quad (19)$$

This is the solid blue line in Fig. 10. As time rolls by, the student becomes increasingly confident that she has, indeed, missed the last bus of the night. Once the seven minute mark has passed, she knows that all is lost.

This is a nice start on a description of the student's state of belief, but it's troubling in part because of that absolute certainty. Are you *sure*, we might ask, that the bus won't come at seven minutes and five seconds? Perhaps you were mistaken about how often the buses run—was it every seven minutes, or every eight minutes? Let's imagine that the student enlarges her theory  $b$  to allow for the possibility that the buses are not spaced every seven minutes, but that the spacing could be anywhere between three minutes and nine (as in “the buses run regularly, every three to nine minutes”). We can add a new parameter to  $b$ ,  $s$ , the spacing, to get we have  $P(t|b, \phi, s)$ , the probability that we're still waiting at time  $t$  if the last bus left  $\phi$  minutes before we arrived and the buses arrive once every  $s$  minutes. With a little thought, we can integrate out the  $\phi$  to get

$$P(t|b, s) = \frac{1}{s}(s - t), \quad (20)$$

where  $s$  is the spacing, and  $P(t|b, s)$  is zero when  $t$  is greater than  $s$ . We now just need a prior on  $s$ , which we'll take to be uniform between three and nine;  $P(s)$  is a constant, in other words, equal to  $1/6$  (by a reasoning process analogous to how we got  $P(\phi)$  equal to  $1/7$  when  $s$  was seven). Let's do the integral

$$P(t|b) = \int_3^9 P(t|b, s)P(s) ds \quad (21)$$

When  $t$  is less than three, any spacing is allowable, and we have

$$P(t|b) = \frac{1}{6} \int_3^9 \frac{1}{s}(s - t) ds = \frac{1}{6} (s - t \log(s)) \Big|_3^9 = \frac{1}{6} (6 - 3 \log 3) \quad (22)$$

When  $t$  is greater than three, but less than nine, some spacings, with  $s$  less than  $t$ , are no longer possible, and so we have to be careful to include only the allowable ones, for which  $P(t|b, s)$  is non-zero,

$$P(t|b) = \frac{1}{6} \int_t^9 \frac{1}{s}(s - t) ds = \frac{1}{6} (s - t \log(s)) \Big|_t^9 = \frac{1}{6} (9 - t - t \log \frac{9}{t}), \quad (23)$$

and when  $t$  is greater than nine,  $P(t|b)$  is zero (we're doomed again—you can't have been waiting that long if the buses were still running). The resulting  $P(a|t)$  that comes from this enlarged theory space is shown as the dashed red line in Fig. 10. Notice that at small waiting times, the possibility that the bus could have come as quickly as three minutes, even if it had just left, makes us nervous: we get more worried that theory  $a$  might be true. That's the red dashed line being higher than the blue solid line. This cross-over depends on exactly what we think about the possible range of arrival gaps.

Note that even in this enlarged theory space, we become completely certain that the buses have stopped at some point—now at nine minutes, not seven. A more realistic theory yet might put non-uniform prior on  $s$ , such that  $P(s)$  was high for “normal” gaps, like three minutes or seven minutes



or nine minutes, and lower (but not zero) for longer ones. One could add in a delay probability, to cover traffic jams or bus drivers getting lost or having to throw drunkards off the bus—left as an exercise to the reader.

One can extract more from these equations. For example, you can think about writing down  $P(\phi|t, b, s)$ , or, in words, what’s the probability that the most recent bus left  $\phi$  minutes before I arrived, given that I’ve been waiting for time  $t$ , and the buses are still coming with frequency  $s$  (theory  $b$  is true). You can combine this with priors on  $s$  to answer questions like “what’s the probability that the bus will arrive in the next  $t'$  minutes, if the buses are still running”:

$$P(t') = \int_3^9 P(\phi = s - (t + t')|t, b, s)P(s) ds. \quad (24)$$

Take some time to work through this equation. In words, the formula in the integrand means “what’s the probability (degree of belief) that  $\phi + t + t'$  is equal to the spacing  $s$ , given that I’ve been waiting for time  $t$ ”, which is then weighted by the prior belief in that spacing,  $P(s)$ . Once you get handy with conditional probabilities and Bayes rule, it’s “simple” to write down answers to questions like these.

## 11 Acknowledgements

This article is in draft form and modified from text from the 2016 Santa Fe Institute Complex Systems Summer School [http://bit.ly/csss\\_bib](http://bit.ly/csss_bib), Indiana University’s seminar on Large-Scale Social Systems, I400/H400/I590, <http://bit.ly/lssp2014>, and Ani Patel’s upper-level class in Cognitive Science at Tufts University in 2016. I thank Artemy Kolchinsky (IU), Austin Hart (American), Drew Cabaniss (UNC), Charley Lineweaver (ANU), Daniel Dennett (Tufts), Sam Scarpino (University of Vermont), Marion Dumas (Santa Fe Institute), Joel Erickson (Two Sigma), Michael Mauboussin (Columbia Business School), Alje van Dam (Public Health Service Amsterdam), Jeffrey R. Chasnov (Hong Kong University of Science and Technology), the forty-two students in Ani Patel’s class, and readers from the general public including Phil Moyer and Nick Noel, for detailed feedback.

## References

- [1] E. T. Jaynes. Probability theory with applications in science and engineering. 1954. Available at <http://bayes.wustl.edu/etj/science.pdf.html>.
- [2] John R Searle. *Rationality in action*. MIT press, Cambridge, MA, 2003.
- [3] Ole Peters. The time resolution of the St Petersburg paradox. *Phil. Trans. R. Soc. A*, 369(1956):4913–4931, 2011.
- [4] John Maynard Keynes. *A treatise on probability*. Macmillan, London, UK, 1921.
- [5] William Whewell. *The Philosophy of the Inductive Sciences: Founded Upon Their History*, volume 2. John W. Parker, West Strand, London, United Kingdom, 2nd edition, 1847. Pages 46–74. Digitized online <https://ia600200.us.archive.org/13/items/philosophyinduc02whewgoog/philosophyinduc02whewgoog.pdf>; call number Q175.W43 P4.
- [6] Robert J. Aumann. Agreeing to disagree. *The Annals of Statistics*, 4(6):1236–1239, 1976.
- [7] Scott Aaronson. The complexity of agreement. In *Proceedings of the Thirty-seventh Annual ACM Symposium on Theory of Computing*, STOC ’05, pages 634–643, New York, NY, USA, 2005. ACM. See also Scott’s discussion at <http://www.scottaaronson.com/blog/?p=2410>.
- [8] Eric S. Raymond. *The New Hacker’s Dictionary*. MIT Press, Cambridge, MA, USA, 3rd edition, 1996. See <http://catb.org/jargon/html/koans.html>.
- [9] Samuel S-H Wang. Origins of presidential poll aggregation: A perspective from 2004 to 2012. *International Journal of Forecasting*, 31(3):898–909, 2015.