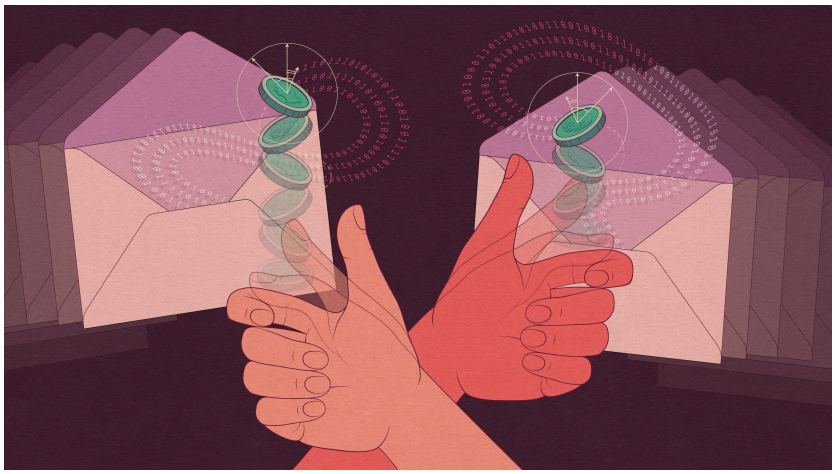
Quantamagazine

How Shannon Entropy Imposes Fundamental Limits on Communication

By **Kevin Hartnett**

September 6, 2022

What's a message, really? Claude Shannon recognized that the elemental ingredient is surprise.



To communicate a series of random events, such as coin flips, you need to use a lot of information, since there's no structure to the message. Shannon entropy measures this fundamental constraint.

Kristina Armitage/Quanta Magazine

If someone tells you a fact you already know, they've essentially told you nothing at all. Whereas if they impart a secret, it's fair to say something has really been communicated.

This distinction is at the heart of Claude Shannon's theory of information. Introduced in an epochal 1948 paper, "A Mathematical Theory of Communication," it provides a rigorous mathematical framework for quantifying the amount of information needed to accurately send and receive a message, as determined by the degree of uncertainty around what the intended message could be saying.

Which is to say, it's time for an example.

In one scenario, I have a trick coin — it's heads on both sides. I'm going to flip it twice. How much information does it take to communicate the result? None at all, because prior to receiving the message, you have complete certainty that both flips will come up heads.

In the second scenario I do my two flips with a normal coin — heads on one side, tails on the other. We can communicate the result using binary code: 0 for heads, 1 for tails. There are four possible messages — 00, 11, 01, 10 — and each requires two bits of information.

So, what's the point? In the first scenario you had complete certainty about the contents of the message, and it took zero bits to transmit it. In the second you had a 1-in-4 chance of guessing the right answer — 25% certainty — and the message needed two bits of information to resolve that ambiguity. More generally, the less you know about what the message will say, the more information it takes to convey.

Shannon was the first person to make this relationship mathematically precise. He captured it in a formula that calculates the minimum number of bits — a threshold later called the Shannon entropy — required to communicate a message. He also showed that if a sender uses fewer bits than the minimum, the message will inevitably get distorted.

"He had this great intuition that information is maximized when you're most surprised about learning about something," said Tara Javidi, an information theorist at the University of California, San Diego.

The term "entropy" is borrowed from physics, where entropy is a measure of disorder. A cloud has higher entropy than an ice cube, since a cloud allows for many more ways to arrange water molecules than a cube's crystalline structure does. In an analogous way, a random message has a high Shannon entropy — there are so many possibilities for how its information can be arranged — whereas one that obeys a strict pattern has low entropy. There are also formal similarities in the way that entropy is calculated in both physics and information theory. In physics, the formula for entropy involves taking a logarithm of possible physical states. In information theory, it's the logarithm of possible event outcomes.

The logarithmic formula for Shannon entropy belies the simplicity of what it captures — because another way to think about Shannon entropy is as the number of yes-or-no questions needed, on average, to ascertain the content of a message.

For instance, imagine two weather stations, one in San Diego, the other in St. Louis. Each wants to send the seven-day forecast for its city to the other. San Diego is almost always sunny, meaning you have high confidence about what the forecast will say. The weather in St. Louis is more uncertain — the chance of a sunny day is closer to 50-50.

Claude Shannon at Bell Labs in 1954.

Estate of Francis Bello/Science Source

How many yes-or-no questions would it take to transmit each seven-day forecast? For San Diego, a profitable first question might be: Are all seven days of the forecast sunny? If the answer is yes (and there's a decent chance it will be), you've determined the entire forecast in a single question. But

with St. Louis you almost have to work your way through the forecast one day at a time: Is the first day sunny? What about the second?

The more certainty there is around the content of a message, the fewer yes-or-no questions you'll need, on average, to determine it.

To take another example, consider two versions of an alphabet game. In the first, I've selected a letter at random from the English alphabet and I want you to guess it. If you use the best possible guessing strategy, it will take you on average 4.7 questions to get it. (A useful first question would be, "Is the letter in the first half of the alphabet?")

In the second version of the game, instead of guessing the value of random letters, you're trying to guess letters in actual English words. Now you can tailor your guessing to take advantage of the fact that some letters appear more often than others ("Is it a vowel?") and that knowing the value of one letter helps you guess the value of the next (q is almost always followed by u). Shannon calculated that the entropy of the English language is 2.62 bits per letter (or 2.62 yes-or-no questions), far less than the 4.7 you'd need if each letter appeared randomly. Put another way, patterns reduce uncertainty, which makes it possible to communicate a lot using relatively little information.

Note that in examples such as these, you can ask better or worse questions. Shannon entropy sets an inviolable floor: It's the absolute minimum number of bits, or yes-or-no questions, needed to convey a message.

"Shannon showed there is something like the speed of light, a fundamental limit," said Javidi. "He showed that Shannon entropy is a fundamental limit for how much we can compress a source, without risking distortion or loss."

Today, Shannon entropy serves as a yardstick in many applied settings, including information compression technology. That you can zip a large movie file, for example, owes to the fact that pixel colors have a statistical pattern, the way English words do. Engineers can build probabilistic models for patterns of pixel colors from one frame to the next. The models make it possible to calculate the Shannon entropy by assigning weights to patterns and then taking the logarithm of the weight for all the possible ways pixels could appear. That value tells you the limit of "lossless" compression — the absolute most the movie can be compressed before you start to lose information about its contents.

Any compression algorithm's performance can be compared to this limit. If you're far from it, you have an incentive to work harder to find a better algorithm. But if you're close to it, you know that the information laws of the universe prevent you from doing much better.