

Is Information Digital?

Why CPS is fundamentally more than computation

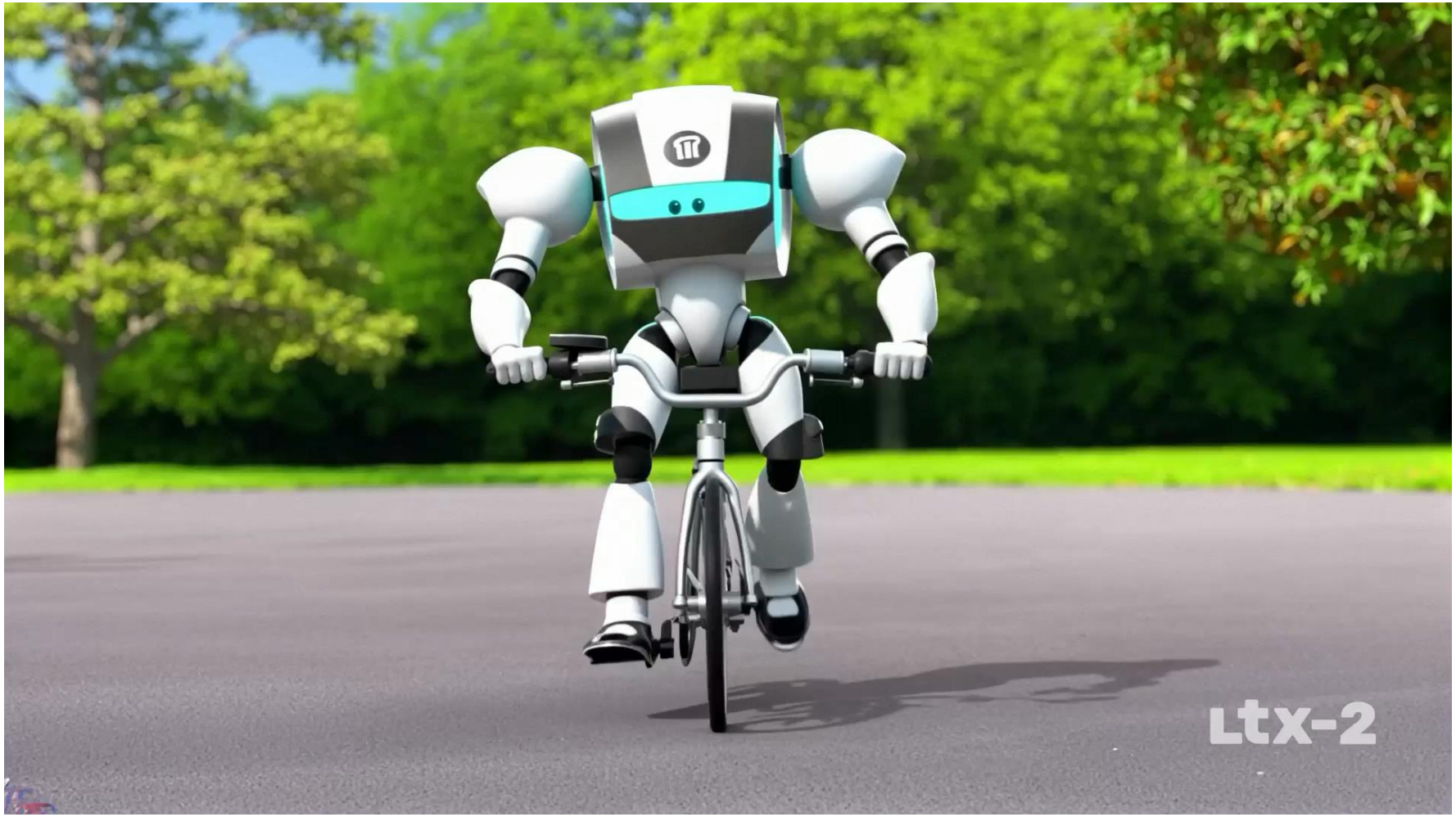
Edward A. Lee

UC Berkeley

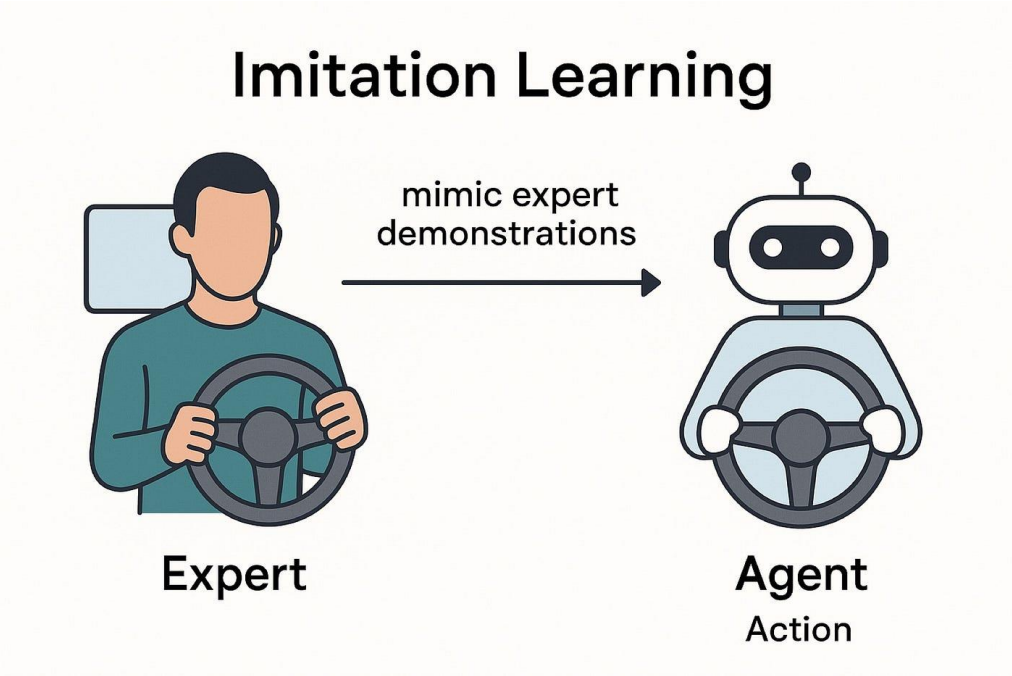
eal@berkeley.edu

Data-Driven Design:
Can you learn to ride a bicycle by watching others ride?

Video generated by LTX Studio by Lightricks (2026)



Practical Question: Will Imitation Learning Work?



Sebastian Buzdugan
July 2025



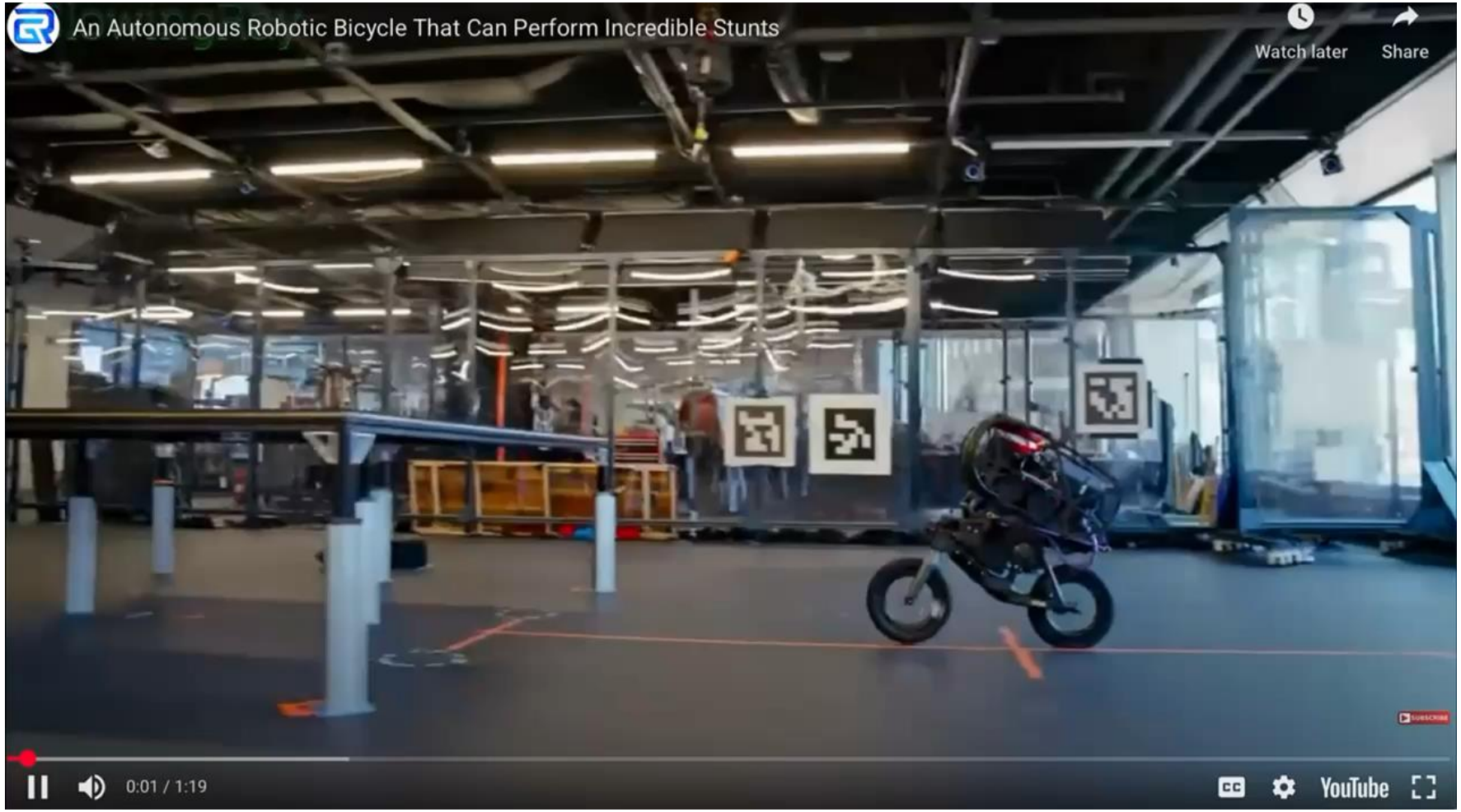
NVIDIA, Feb. 2026

Can more data overcome the gap?

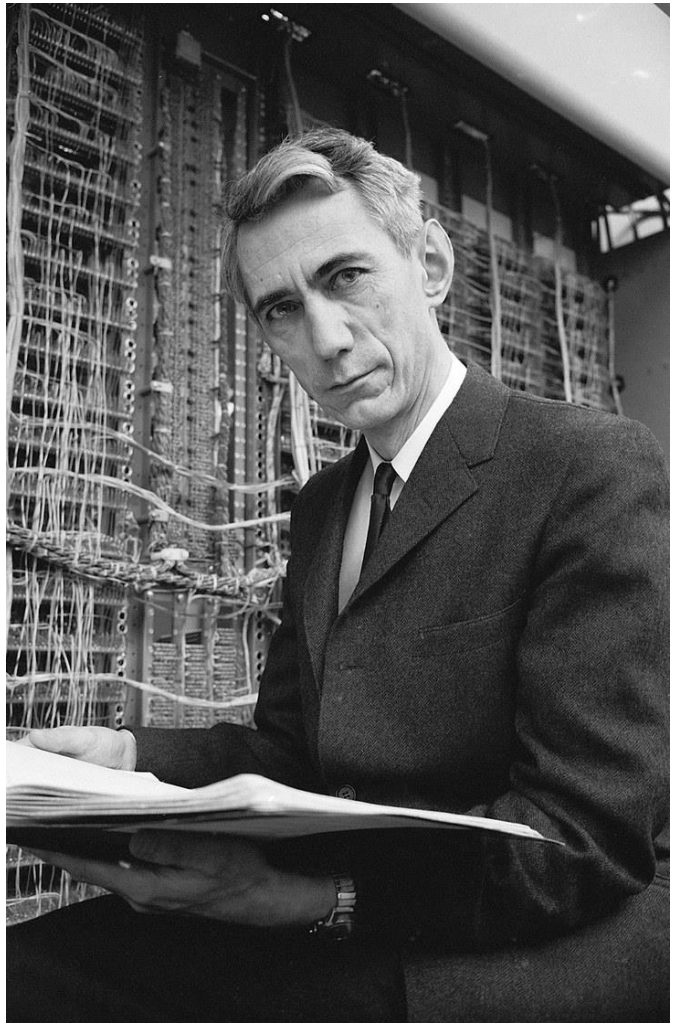
Machines learn to navigate the physical world by *being* it, not by *watching* it.

Ultra Mobility Vehicle (UMV) from the Robotics and AI Institute (RAI).

Video by Glowing Ray (2025)



The Meaning of Information



The Bell System Technical Journal

Vol. XXVII *July, 1948* *No. 3*

A Mathematical Theory of Communication

By C. E. SHANNON

Introduction

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. As was pointed out by Hartley the most natural choice is the logarithmic function. Although this definition must be generalized considerably when we consider the influence of the statistics of the message and when we have a continuous range of messages, we will in all cases use an essentially logarithmic measure.

The logarithmic measure is more convenient for various reasons:

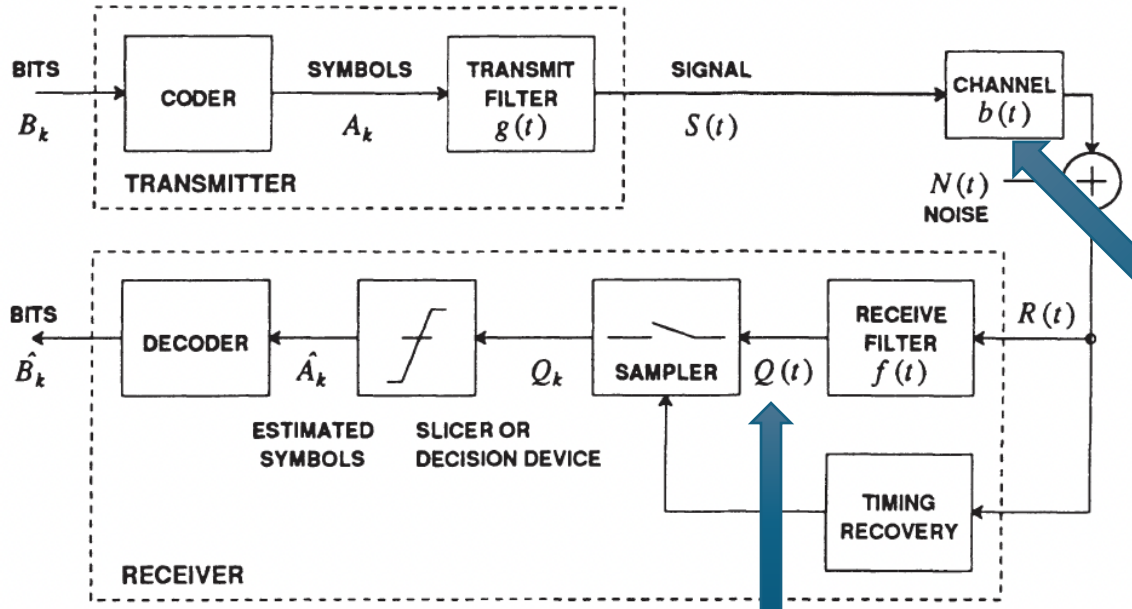
1. It is practically more useful. Parameters of engineering importance

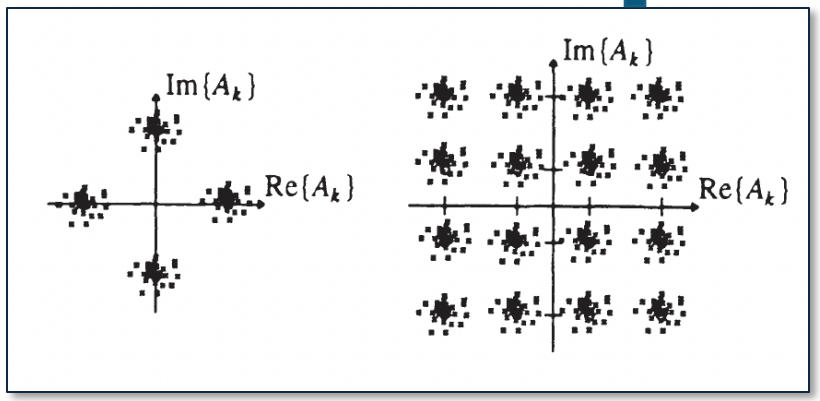
[1] Nyquist, H., "Certain Factors Affecting Telegraph Speed," *Bell System Technical Journal*, April 1924, p. 324; "Certain Topics in Telegraph Transmission Theory," *A. I. E. E. Trans.*, v. 47, April 1928, p. 617.

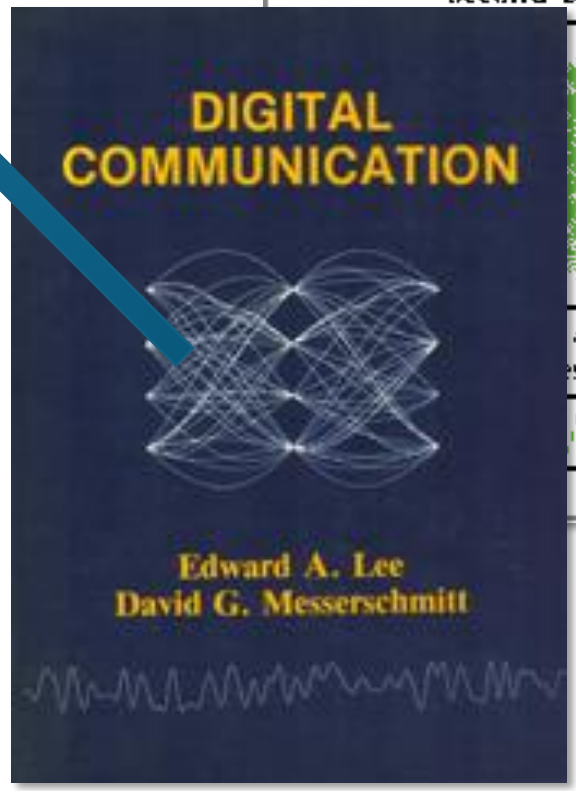
[2] Hartley, R. V. L., "Transmission of Information," *Bell System Technical Journal*, July 1928, p. 535.

379

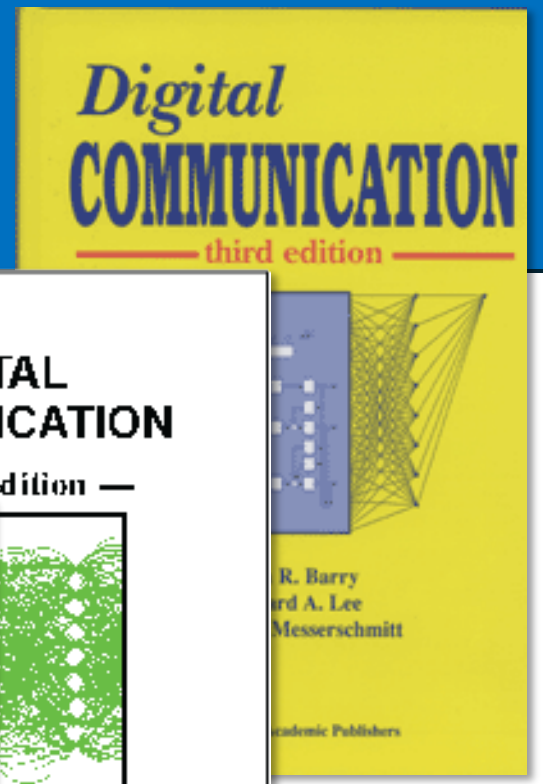
Shannon's Context: Communication Over a Noisy Channel







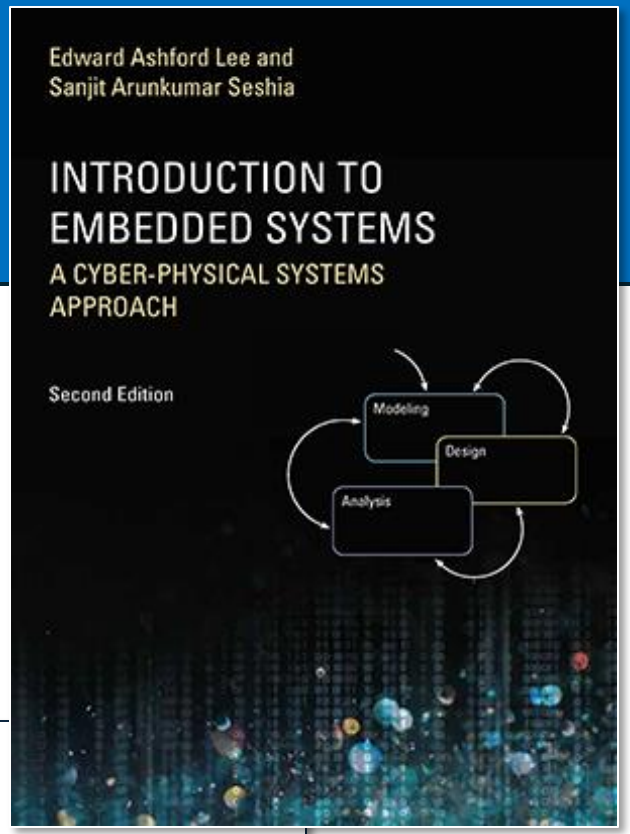




Kluwer
1988, 1994, 2004

Measurement of a Physical Phenomenon…

… is communication
over a
noisy channel.



MIT Press
2017

Epistemology

Ontology

7.1.5 Noise

By definition, **noise** is the part of a signal that we do not want. If we want to measure $x(t)$ at time t, but we actually measure $x'(t)$), then the noise is the difference,

$$n(t) = x'(t) - x(t).$$

Equivalently, the actual measurement is

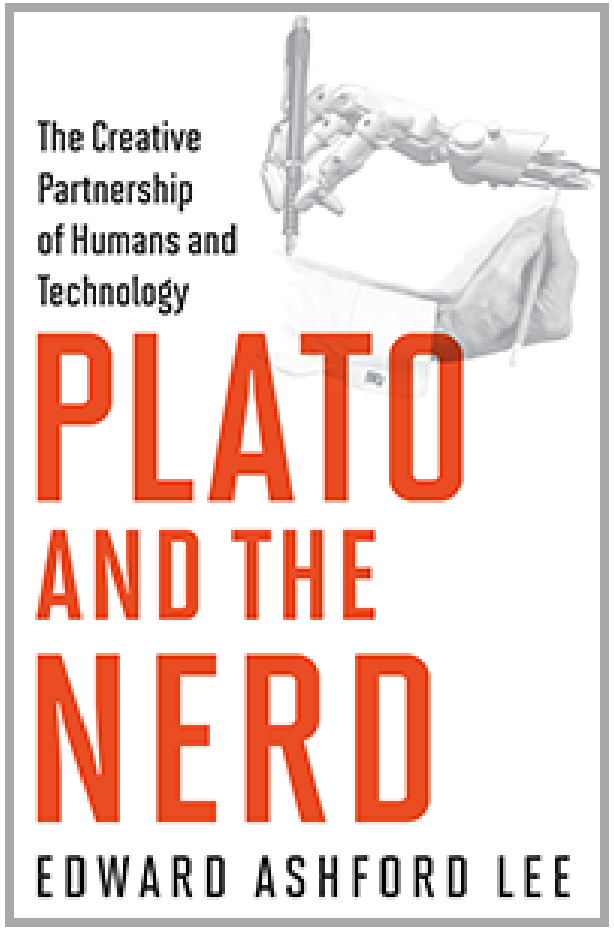
$$x'(t) = x(t) + n(t), \tag{7.5}$$

a sum of what we want plus the noise.

Shannon Channel Capacity Theorem

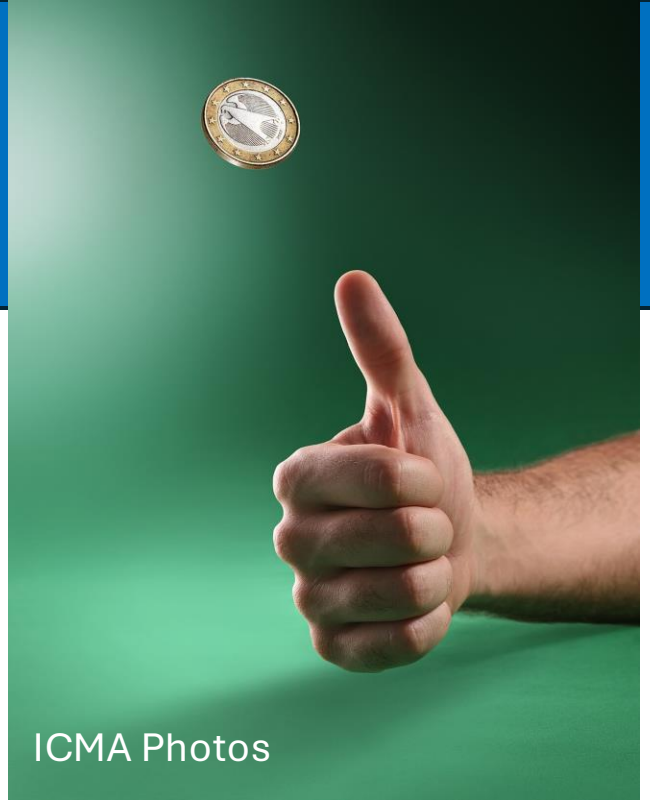
Any noisy measurement yields a finite number of bits of information.

But this does not mean that the physical object contains only a finite number of bits of information.



MIT Press
2017

Information as Surprise



Consider an unfair coin, $p_H = 0.9$, $p_T = 0.1$

Information gained from a trial:

$-\log_2(p_H) \approx 0.15$ bits

$-\log_2(p_T) \approx 3.32$ bits

Tails is more unlikely, more surprising, more information.

This fits well with Bayesian interpretation of probability.

Bayesian Probability

In the Bayesian interpretation, a probability is a measure of what we don't know, not a measure of some underlying randomness.

Hence, a higher probability outcome carries less information (we expect it to occur), whereas observing a lower probability carries more information (we do not expect it).

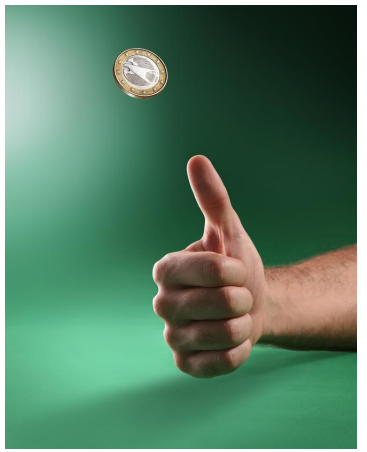


Portrait of Reverend Thomas Bayes (1701 - 1761) that is probably not actually him.

Expected Information (Entropy)

$-\log_2(p_H) \approx 0.15$
$-\log_2(p_T) \approx 3.32$

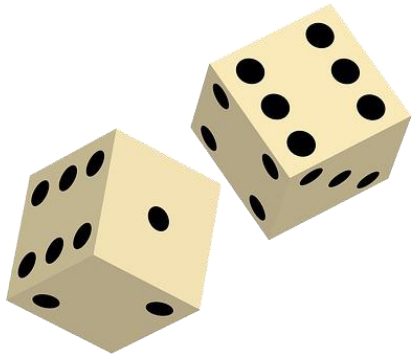
The expected information is the average information from each trial weighted by its probability:



$$H = -0.9\log_2(0.9) - 0.1\log_2(0.1) \approx 0.47 \text{ bits}$$

This is called **entropy** and generalizes to any discrete random variable:

$$H(X) = -\sum_{x \in \Omega} p(x)\log_2(p(x))$$



Relation to Physical Entropy

Consider particles in a gas

$$S = -k_B \sum_{x \in \Omega} p(x) \ln(p(x))$$

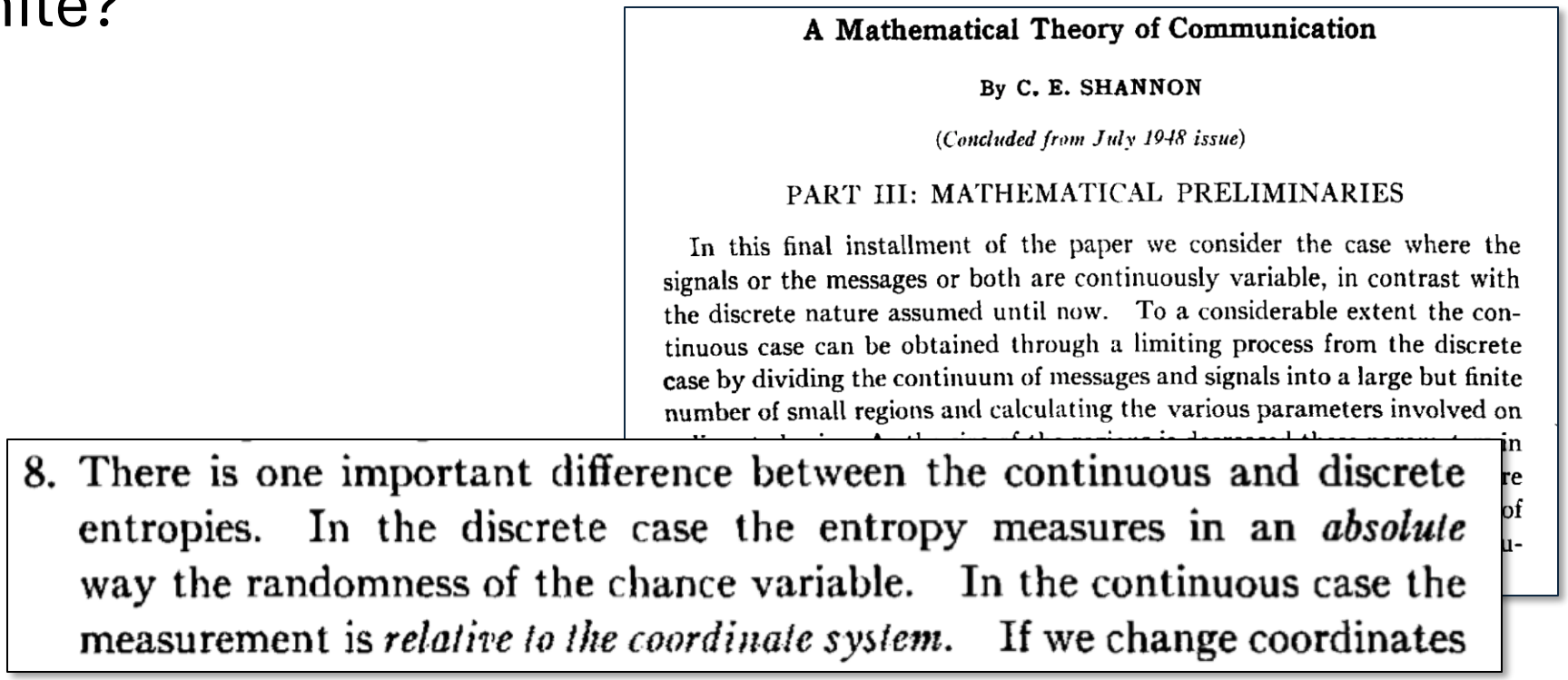
Here, Ω is the set of all possible states of the system that are compatible with its macroscopic properties (like temperature and pressure for a gas), and $p(x)$ is the probability of state x occurring.

- Absolute zero => zero entropy
- Equilibrium => maximum entropy

But Wait…

Number of possible states? Isn't this infinite? A continuum?

Shannon's communication problem: number of possible observations? Again, infinite?



A Mathematical Theory of Communication

By C. E. SHANNON

(Concluded from July 1948 issue)

PART III: MATHEMATICAL PRELIMINARIES

In this final installment of the paper we consider the case where the signals or the messages or both are continuously variable, in contrast with the discrete nature assumed until now. To a considerable extent the continuous case can be obtained through a limiting process from the discrete case by dividing the continuum of messages and signals into a large but finite number of small regions and calculating the various parameters involved on

8. There is one important difference between the continuous and discrete entropies. In the discrete case the entropy measures in an *absolute* way the randomness of the chance variable. In the continuous case the measurement is *relative to the coordinate system*. If we change coordinates

Entropy of a Continuous Random Variable

$$H(X) = -\int_{\Omega} f(x)\log_2(f(x))dx$$

vs. $$H(X) = -\sum_{x\in\Omega} p(x)\log_2(p(x))$$

It is no longer correct to interpret this entropy as giving the average number of bits required to encode an outcome!

In fact, with probability one, every outcome will require an ***infinite*** number of bits to encode.

What does this have to do with black holes?



PHYSICAL REVIEW D VOLUME 7, NUMBER 8 15 APRIL 1973

Black Holes and Entropy*

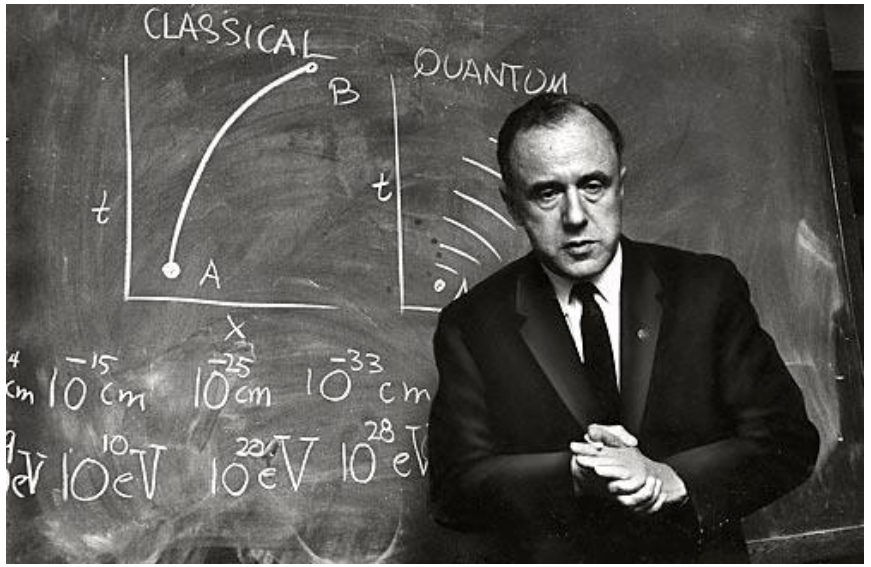
Jacob D. Bekenstein†
Joseph Henry Laboratories, Princeton University, Princeton, New Jersey 08540
and Center for Relativity Theory, The University of Texas at Austin, Austin, Texas 78712‡
(Received 2 November 1972)

There are a number of similarities between black-hole physics and thermodynamics. Most striking is the similarity in the behaviors of black-hole area and of entropy: Both quantities tend to increase irrev
a thermodynamic approach to bla
the theory of information, we dis
tion theory. We show that it is n
measure of information about a b
observer. Considerations of sim
cate that the black-hole entropy
the Planck length times a dimens
making use of the specific prope
theory leads to the same conclus
physical content of the concept o
version of the second law: When
entropy in the black-hole exteric
of this version of the second law
well as by several examples.

straightforward manner. However, our value $\eta = \frac{1}{2}\ln 2$ might presumably be challenged on the grounds that it follows from the assumption that the smallest possible radius of a particle is precisely equal to its Compton wavelength whereas the actual radius is not so sharply defined. Nevertheless, it should be clear that if η is not exactly $\frac{1}{2}\ln 2$, then it must be very close to this, probably within a factor of two. This slight uncertainty in the value of η is the price we pay for not giving our problem a full quantum treatment. However, in what follows we shall suppose that $\eta = \frac{1}{2}\ln 2$. Ex-

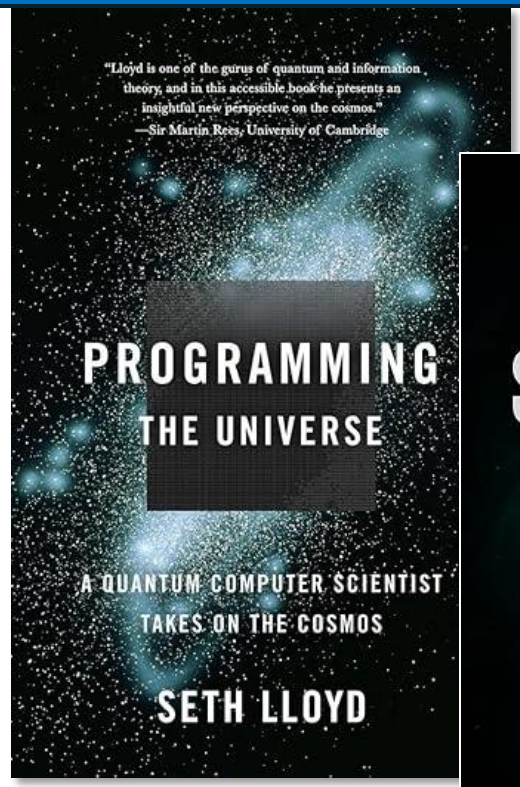
Digital Physics?

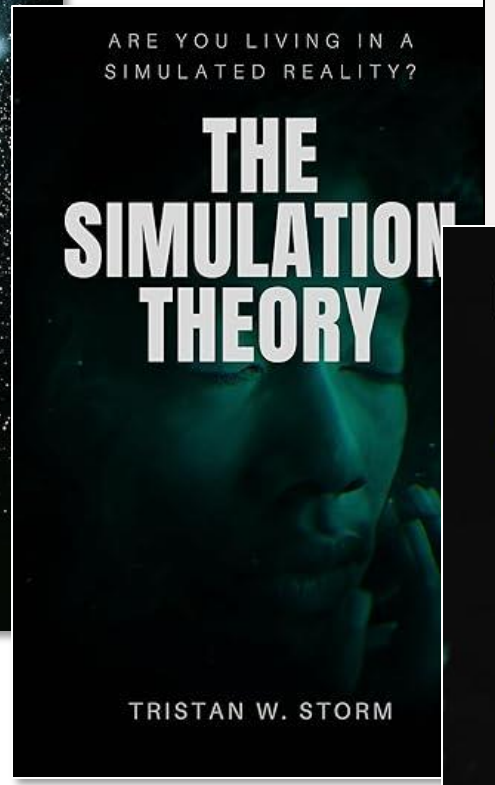
1. The number of possible states of a physical system is finite.
2. Physical processes are digital and algorithmic.
3. Every physical process is a Turing computation.
4. The physical world is a computer.
5. The physical world is a simulation.

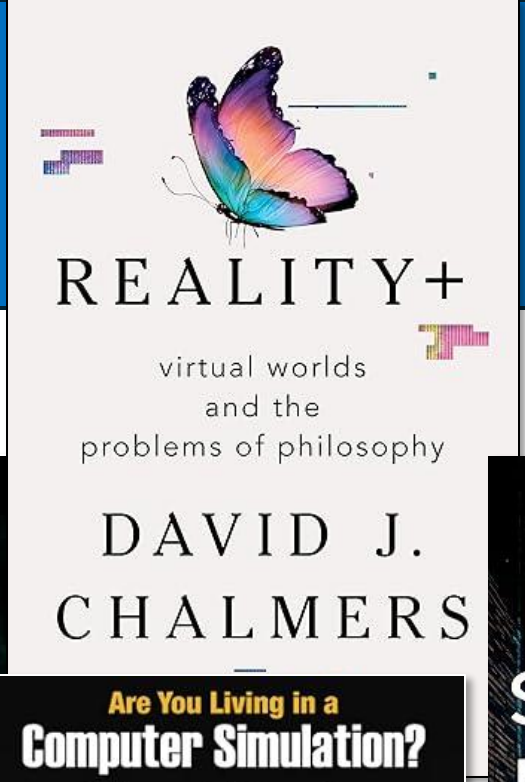


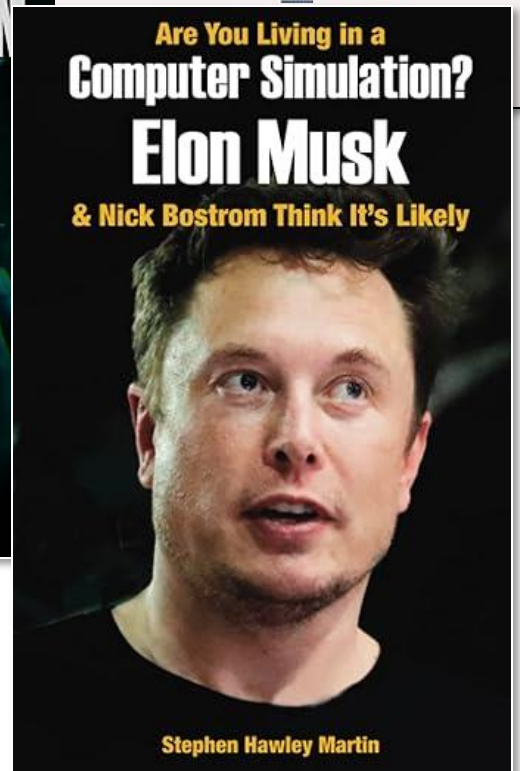
John Archibald Wheeler “It from bit”

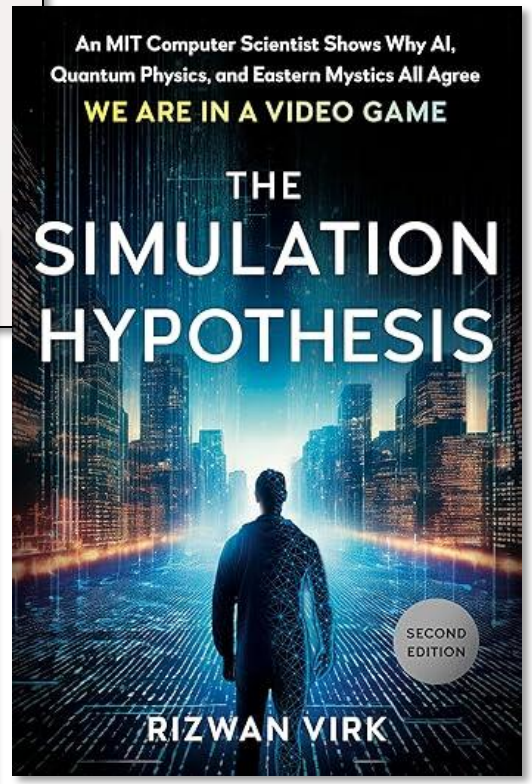
The Simulation Hypothesis











The Mistake

- Digital physics is based on a math error.
- The simulation hypothesis is bunk…

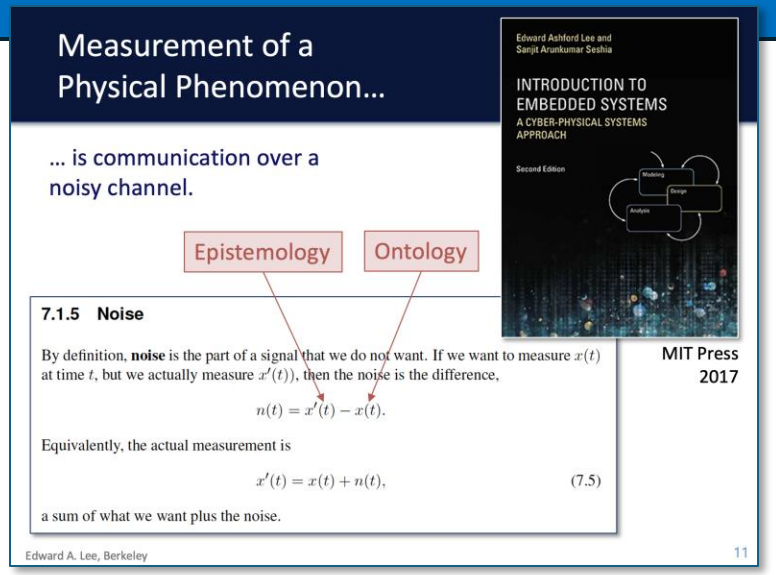
… unless you define computation to be whatever the physical world does, in which case, it's a tautology.

Are we talking about ontology or epistemology?

Mutual Information: Information in a Noisy Measurement

X is a random variable (an unknown).
Y is a noisy measurement of *X*.
Mutual information:

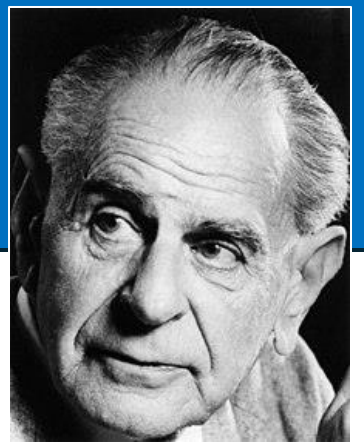
$$I(X;Y) = H(X) - H(X|Y)$$



H(*X*|*Y*) is the entropy remaining in *X* after observing *Y*.

This has units of bits! It tells us that **a noisy measurement never reveals more than a finite number of bits of information about the thing being measured**.

Digital Physics?
Not a Scientific Hypothesis

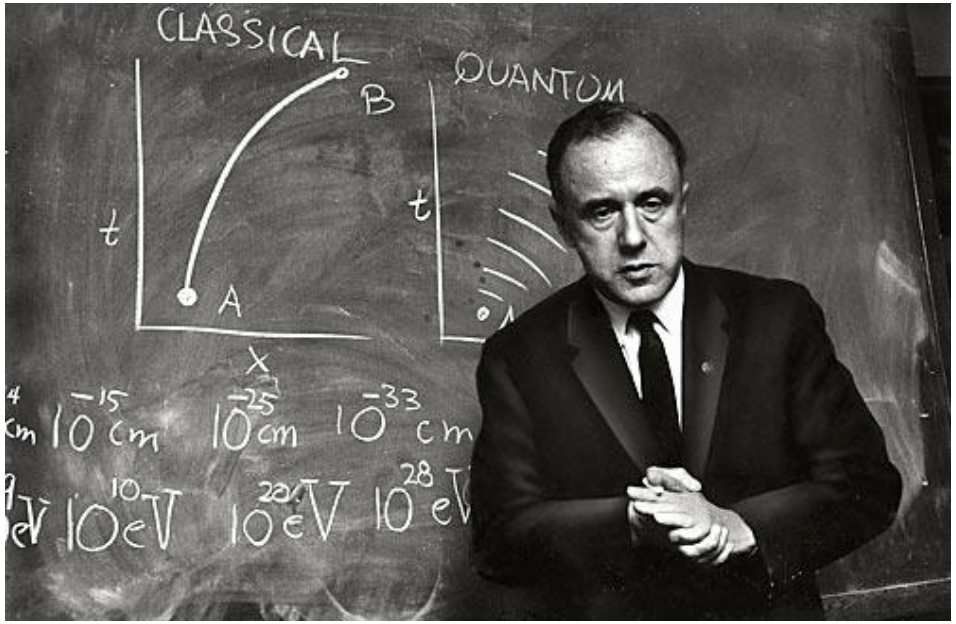


Karl Popper

1. The number of possible states of a physical system is finite.
2. Physical processes are digital and algorithmic.
3. Every physical process is a Turing computation.
4. The physical world is a computer.
5. The physical world is a simulation.

These theses are not falsifiable, and therefor not scientific.

You can only take them on faith.



John Archibald Wheeler “It from bit”

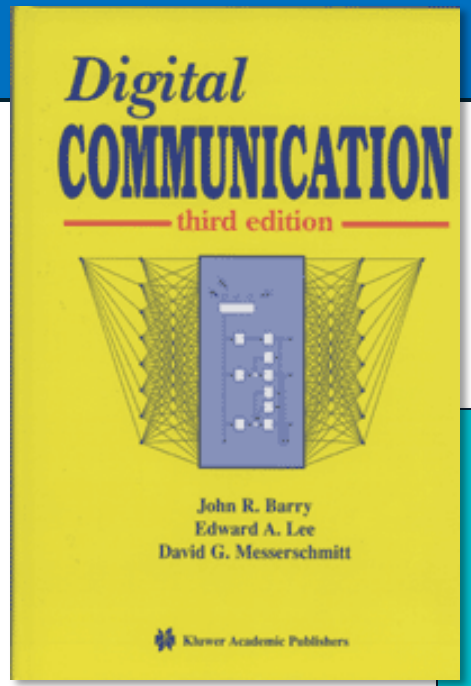
Conclusion

Information theory tells us that there is infinitely more information in reality than in measurements of reality and in computational system.

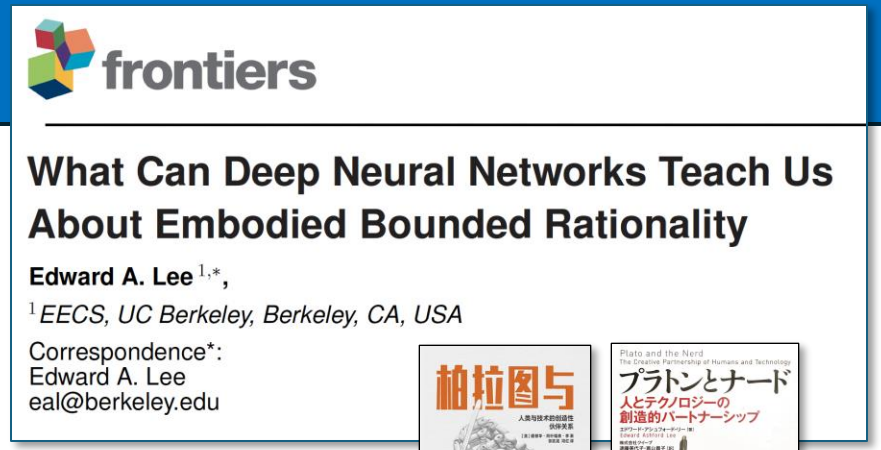
Embodied interaction means becoming part of a physical system (an ontological fact) vs. observation, which is about gaining information about a system by observing it (an epistemic fact).

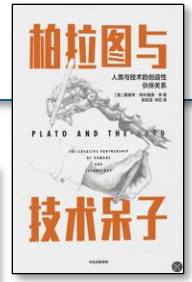
Cyber-physical systems are fundamentally richer than computational systems.

References



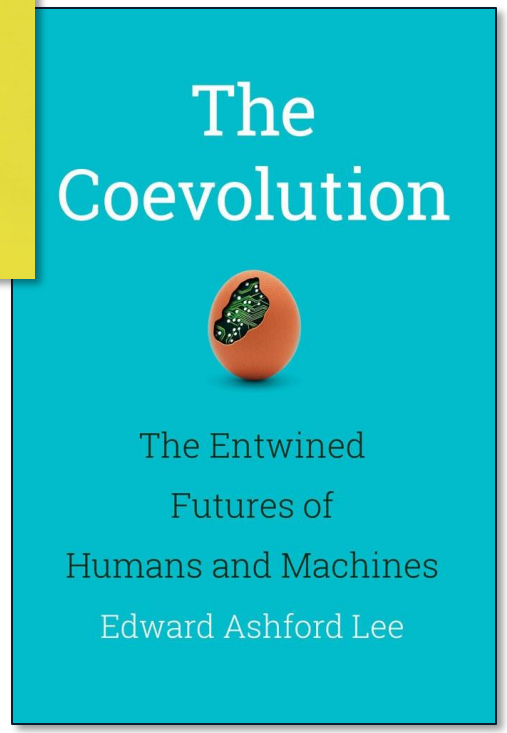
My web page, with links to more references and talks.

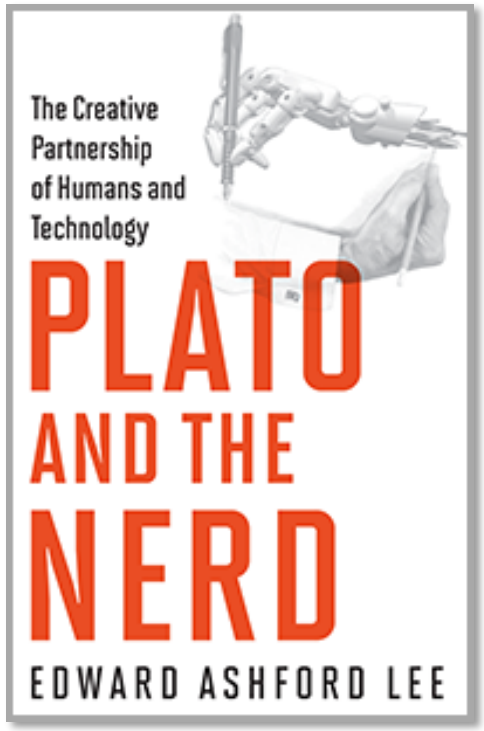


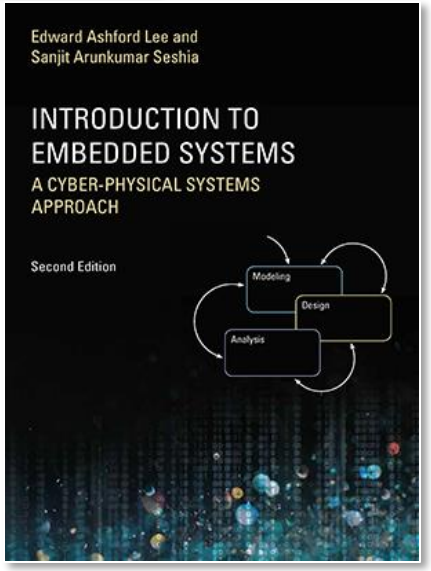












Backup Slides

Information as Surprise

$-\log_2(p_H) \approx 0.15$
$-\log_2(p_T) \approx 3.32$

Why the logarithm? Turns multiplication into addition.

Consider observing two tails in a row, where $p_T = 0.1$:

$$p_{TT} = 0.1 \times 0.1 = 0.01$$

$$-\log_2(p_{TT}) = -\log_2(p_T\, p_T) = -\log_2(p_T) - \log_2(p_T) \approx 6.64$$

Information gained by two trials is the sum of each trial.

Why the minus sign? Makes information positive and increasing with rarity (surprise).

Why the log base two? Makes the units "bits".

Entropy gives a lower bound on the number of bits per trial achievable

$-\log_2(p_H) \approx 0.15$
$-\log_2(p_T) \approx 3.32$

One possible encoding:

HH	0
TH	10
HT	110
TT	111

Expected number of bits per pair of coin tosses:

$$(0.9 \times 0.9) \times 1 + (0.1 \times 0.9) \times 2 + (0.9 \times 0.1) \times 3 + (0.1 \times 0.1) \times 3 = 1.29$$

Expected number of bits per coin toss:

$$1.29/2 = 0.645$$

vs. theoretical limit:

$$H = -0.9 \log_2(0.9) - 0.1 \log_2(0.1) \approx 0.47 \text{ bits}$$

Special cases

Outcome is always the same (zero entropy):

$$p_H = 1, p_T = 0$$

$$H = -p_H \log_2(p_H) - p_T \log_2(p_T) = 0$$

All outcomes equally likely (maximum entropy):

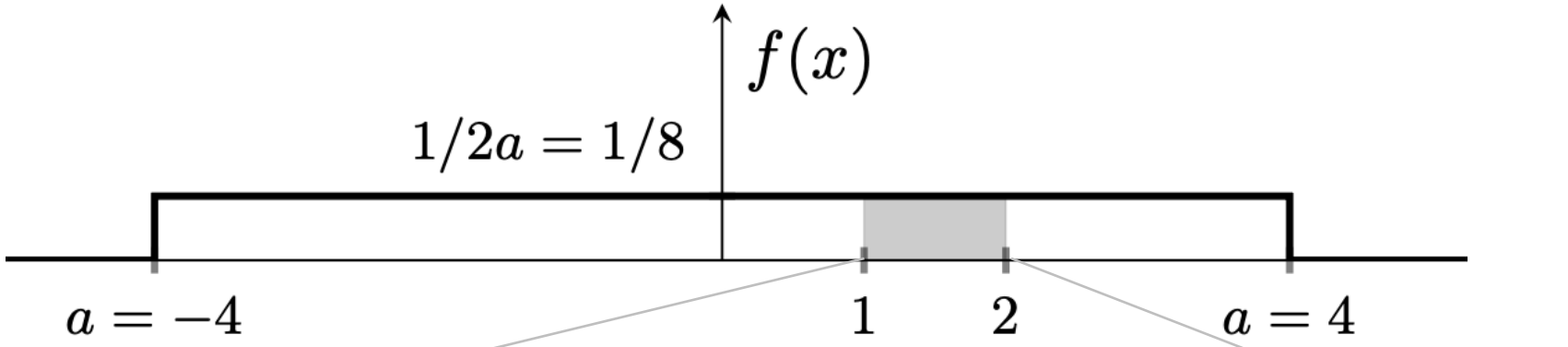
$$p_H = 0.5, p_T = 0.5$$

$$H = -p_H \log_2(p_H) - p_T \log_2(p_T) = 1$$

One bit per toss is required for a fair coin, on average.

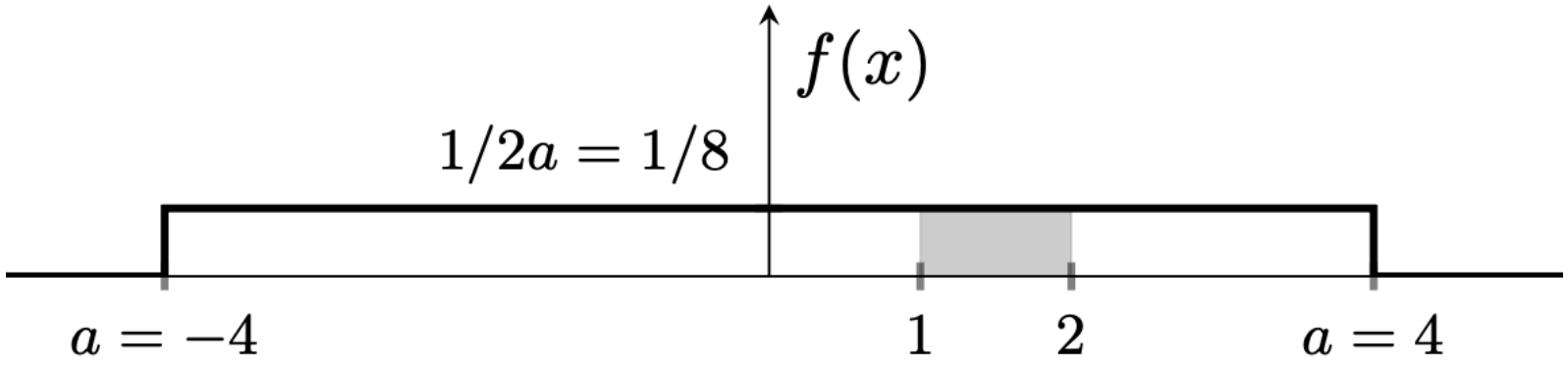
Probability Density

Random variable equally likely to be between –4.0 and 4.0.



Area gives the probability of an outcome occurring in a region.

Entropy of a Uniform Random Variable



$$H(X) = -\int_{\Omega} f(x)\log_2(f(x))dx$$

$$H(X) = -\log_2(1/2a) = \log_2(2a) = 3$$

Can an outcome be encoded with three bits? NO!

Special cases

Outcome is always the same (no uncertainty):

$$f(x) = \delta(x - a) \qquad \text{vs. } p(X = a) = 1$$

$$H(X) = -\infty \qquad \text{vs. } H(X) = 0$$

Maximum entropy distributions:

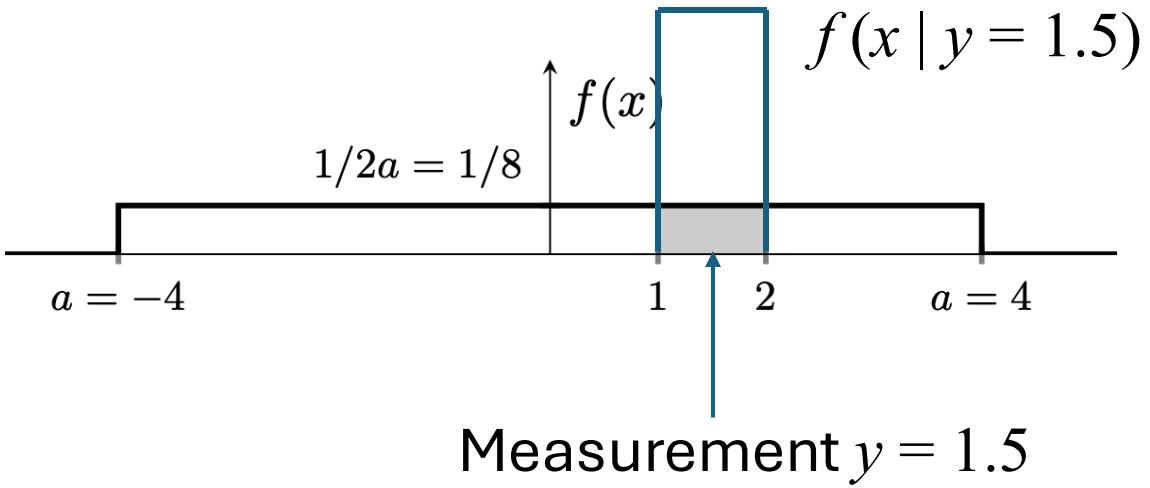
Bounded magnitude: Uniform distribution

Bounded energy (variance): Gaussian distribution

Entropy can be negative. Does not have units of bits!

Mutual Information and Channel Capacity

Suppose the noise is uniformly distributed over $(-0.5, 0.5)$ and we get a measurement $y = 1.5$.



$$H(X \mid y = 1.5) = 0. \quad H(X \mid Y) \cong -0.1.$$

$$I(X;Y) = H(X) - H(X \mid Y) \cong 3 - (-0.1) = 3.1$$

A measurement gives slightly more than 3 bits of information.