Boltzmann's Brain and Wigner's Friend

Charles H. Bennett *(IBM Research)*

Conference in Honor of

Paul Benioff's 85th Birthday 26 May 2016

- Is Dynamics the same as Computation? Argonne, MIT, and IBM in the 1970's and 1980's

- Self-organization, cellular automata and organization as logical depth

- True and False evidence—the Boltzmann Brain problem at equilibrium and in modern cosmology

- Wigner's Friend—what it feels like to be inside an unmeasured superposition

ANL Bldg. 223
ca 1972

Aneesur Rahman
Molecular Dynamics,
RL and Td of Comp,
RADS

Physics of Computation Conference Endicott House MIT May 6-8, 1981

1 Freeman Dyson
2 Gregory Chaitin
3 James Crutchfield
4 Norman Packard
5 Panos Ligomenides
6 Jerome Rothstein
7 Carl Hewitt
8 Norman Hardy
9 Edward Fredkin
10 Tom Toffoli
11 Rolf Landauer
12 John Wheeler

13 Frederick Kantor
14 David Leinweber
15 Konrad Zuse
16 Bernard Zeigler
17 Carl Adam Petri
18 Anatol Holt
19 Roland Vollmar
20 Hans Bremerman
21 Donald Greenspan
22 Markus Buettiker
23 Otto Floberth
24 Robert Lewis

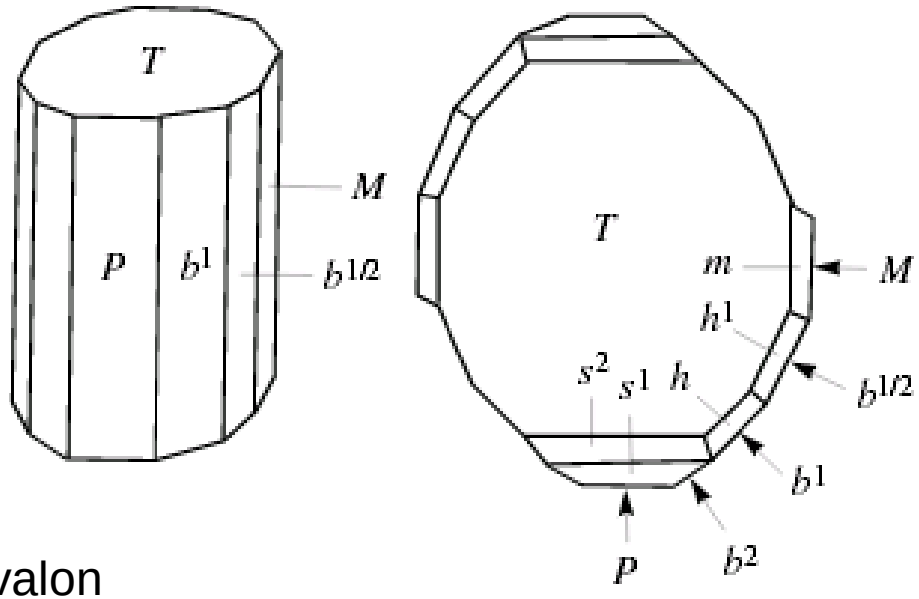
25 Robert Suaya
26 Stan Kugell
27 Bill Gosper
28 Lutz Priese
39 Madhu Gupta
30 Paul Benioff
31 Hans Moravec
32 Ian Richards
33 Marian Pour-El
34 Danny Hillis
35 Arthur Burks
36 John Cocke

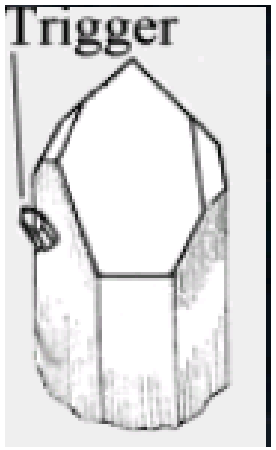
37 George Michaels
38 Richard Feynman
39 Laurie Lingham
40 Thiagarajan
41 ?
42 Gerard Vichniac
43 Leonid Levin
44 Lev Levitin
45 Peter Gacs
46 Dan Greenberger

1984 Cellular Automata Conference in Mosquito Island

Scientific vs. Anthropocentric Thinking

Pasteur's sketch of sodium ammonium tartrate crystal. Chiral location of hemihedral faces e.g. ***h*** is determined by chirality of molecules within.





http://www.neatstuff.net/avalon/texts/Quartz-Configurations.html

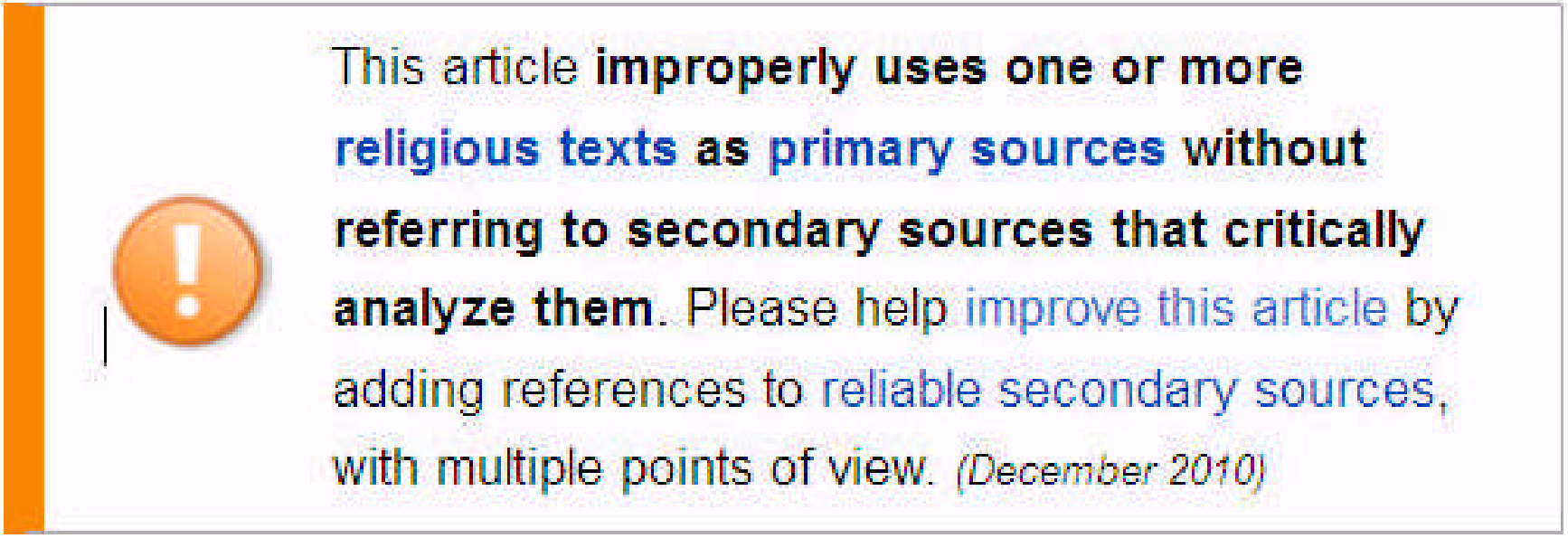
TRIGGER CRYSTALS:
have a **smaller crystal growing out from them**. This `trigger' can be gently squeezed to activate the power of the crystal and strengthen its attributes. These are just used for a surge of a particular kind of energy.

To understand molecules, learn to think like one.

Original form of Occam's Razor:

> "For nothing ought to be posited without a reason given, unless it is self-evident, or known by experience, or proved by the authority of Sacred Scripture"
>
> *William of Ockham (ca.1287 – 1347)*

Scriptures get less respect nowadays



(Wikipedia warning on early version of Mormon Cosmology article)

Can science explain why the world is the way it is, in particular why it is so complicated?

(Some people—perhaps the majority of the human population—think not, and that we should go back to some form of Scripture instead)

But to attack this question in a disciplined fashion, one must first define complexity, the property that increases when a self-organizing system organizes itself.

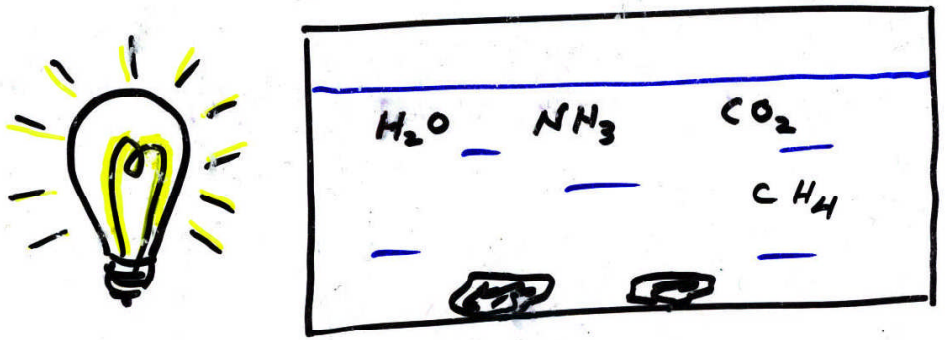
A good scientific theory should give predictions relative to which the phenomena it seeks to explain are typical.

A cartoon by Sidney Harris shows a group of cosmologists pondering an apparent typicality violation

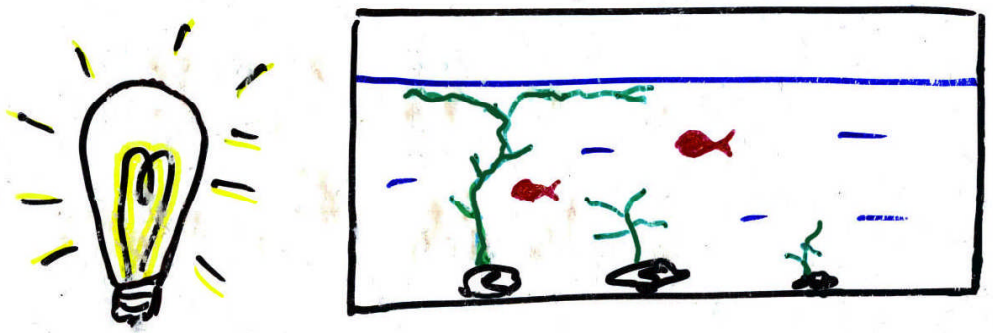
"Now if we run our picture of the universe backwards several billion years, we get an object resembling Donald Duck. There is obviously a fallacy here."

(This cartoon is not too far from problems that actually come up in current cosmology)

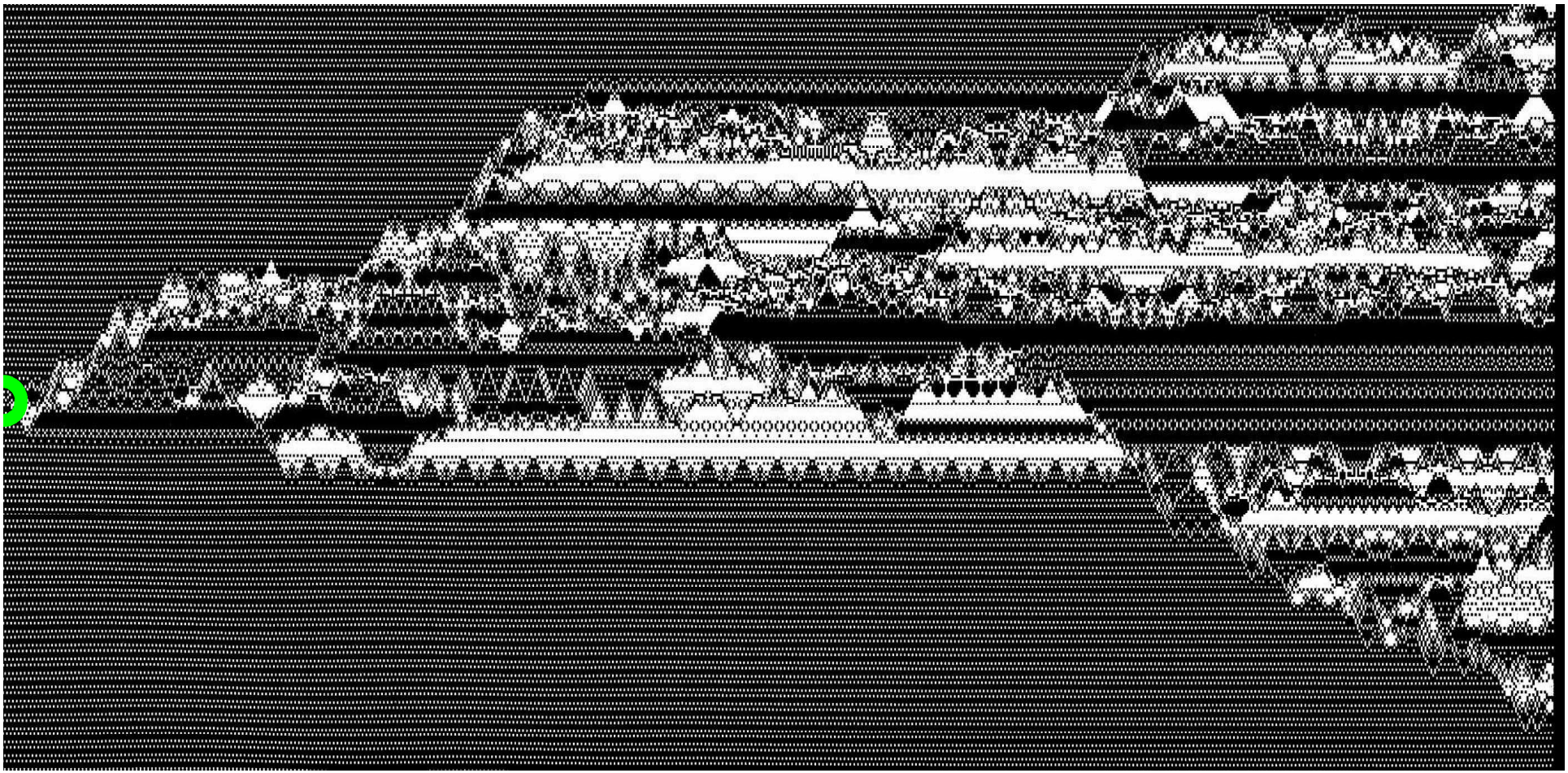
A simple cause can have a complicated effect, but not right away.



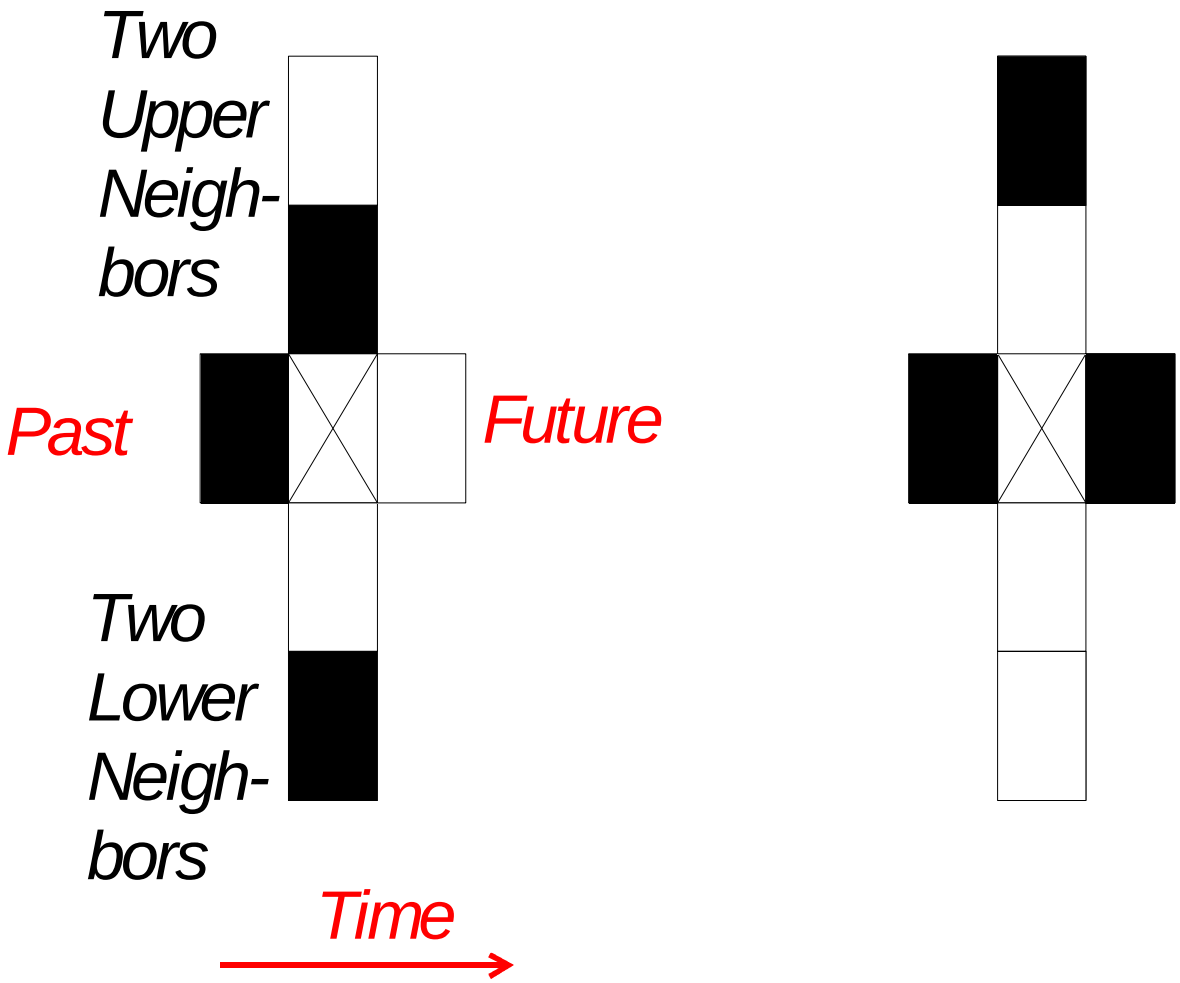
Much later



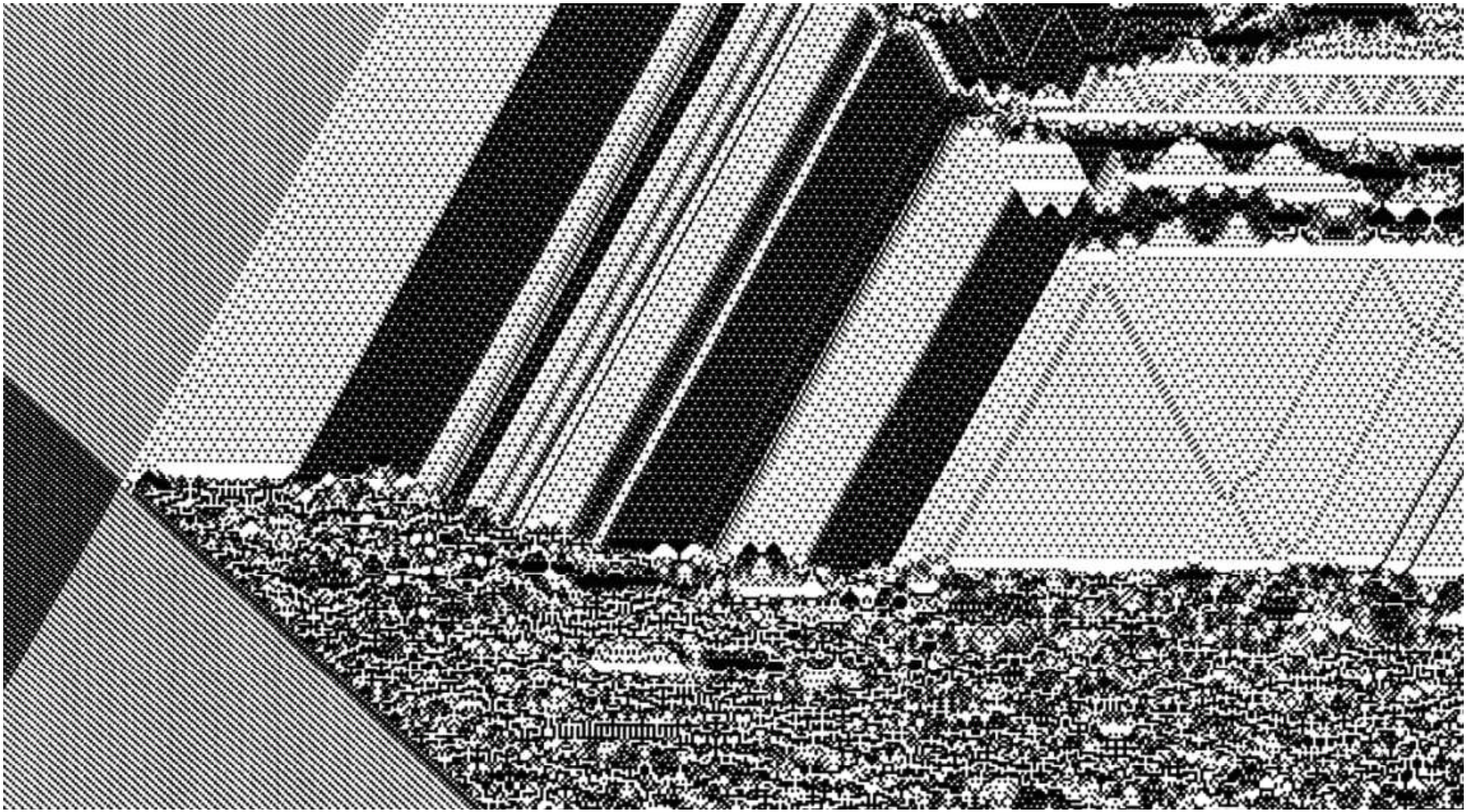
Simple dynamical processes (such as this 1 dimensional reversible cellular automaton) are easier to analyze and can produce structures of growing "complexity" from simple initial conditions. time →



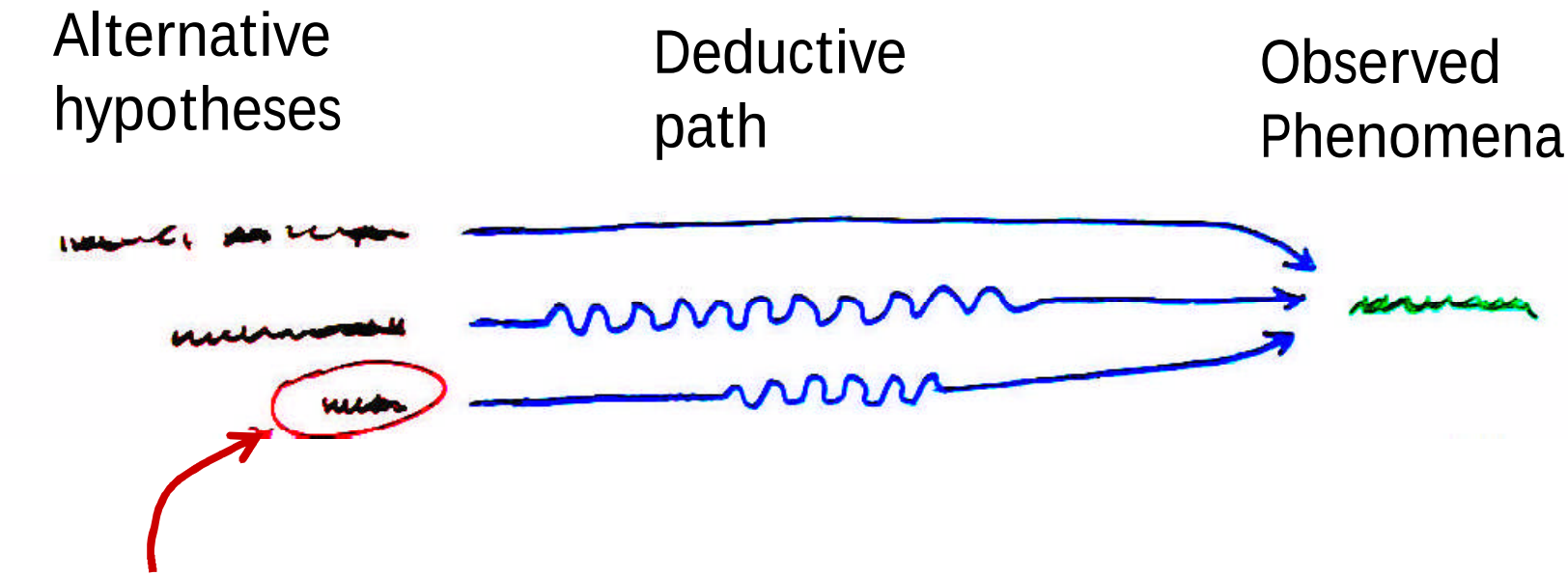
Small irregularity (green) in otherwise periodic initial condition produces a complex deterministic wake.



Range-2, deterministic, 1-dimensional Ising rule. Future differs from past if exactly two of the four nearest upper and lower neighbors are black and two are white at the present time.



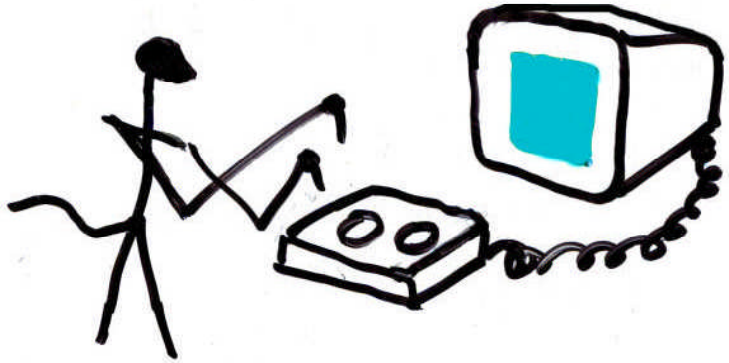
Occam's Razor



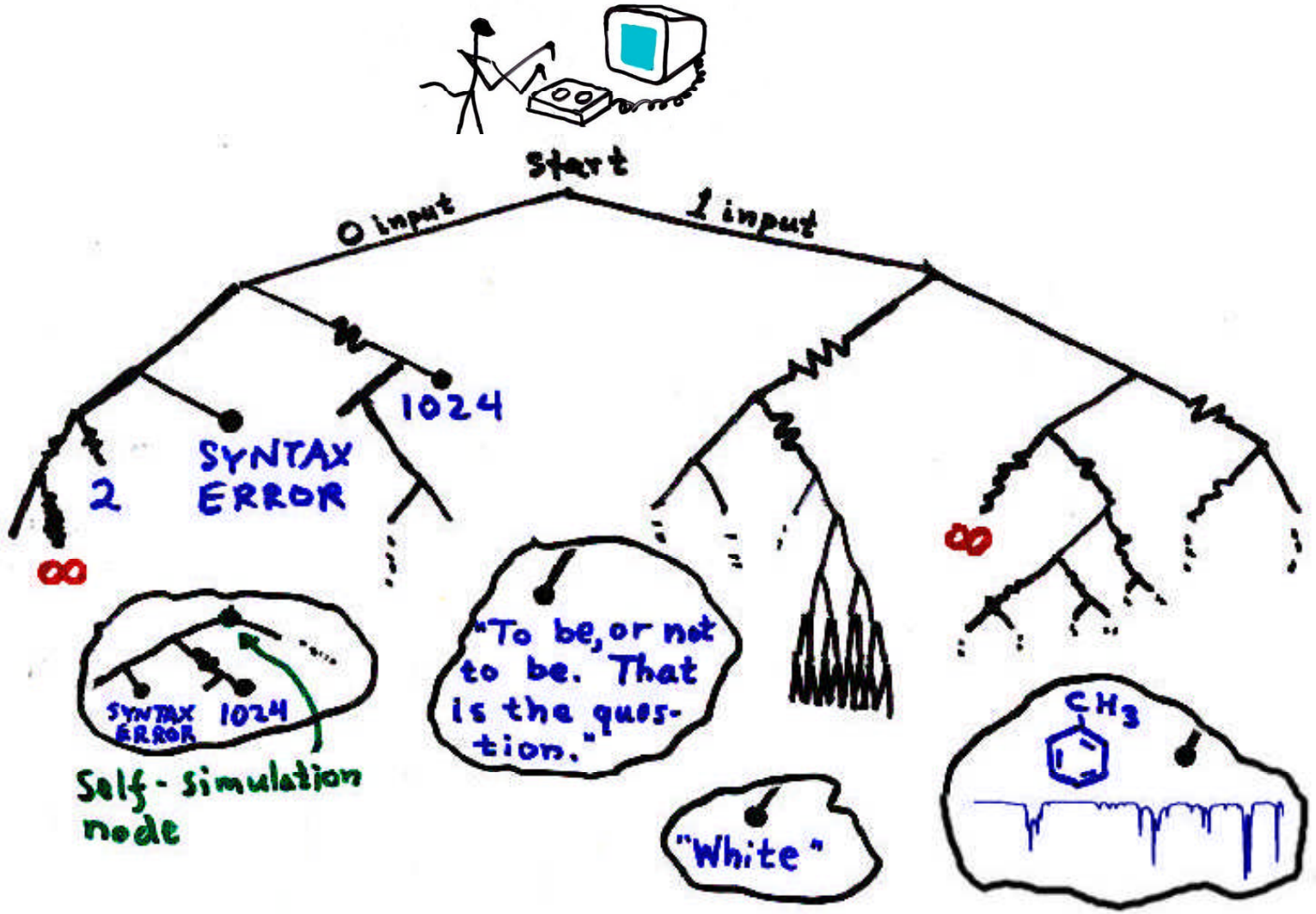
The most economical hypothesis is to be preferred, even if the deductive path connecting it to the phenomena it explains is long and complicated.

But how does one compare economy of hypotheses in a disinterested way?

Algorithmic information, devised in the 1960's by Solomonoff, Kolmogorov, and Chaitin, uses a computerized version of the old idea of a monkey at a typewriter eventually typing the works of Shakespeare.

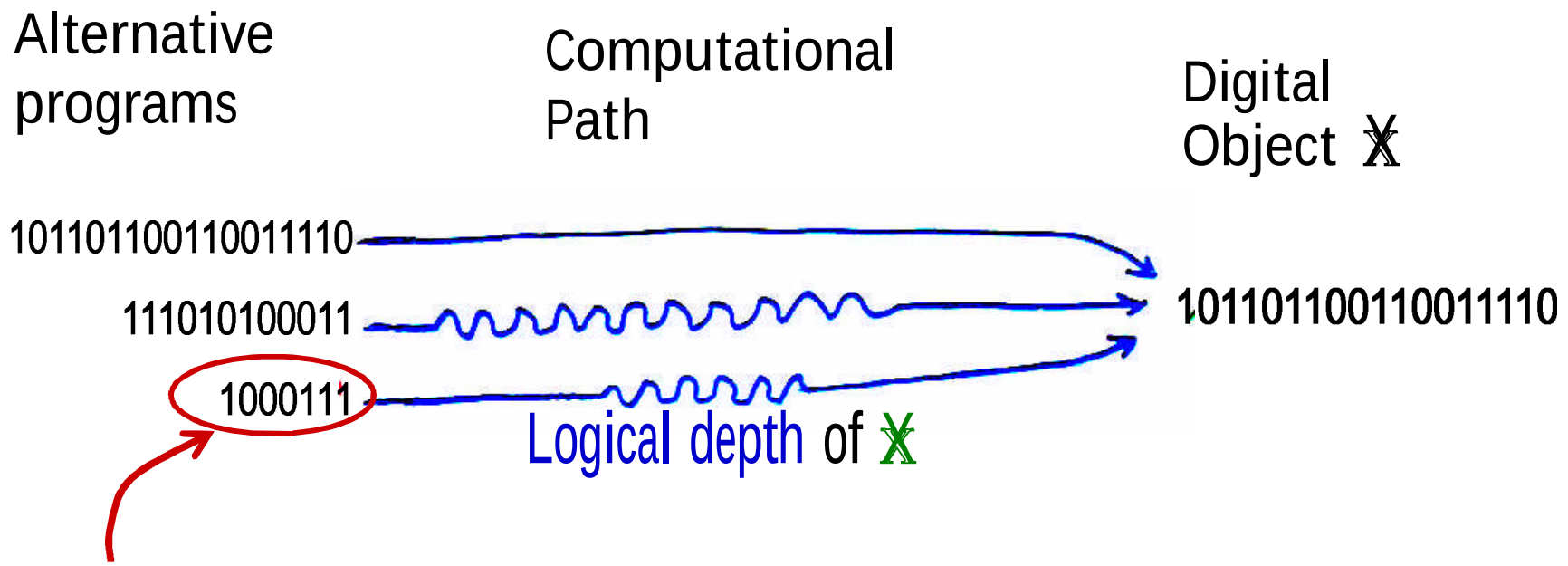


A monkey randomly typing 0s and 1s into a universal binary computer has some chance of getting it to do any computation, produce any output.



This tree of all possible computations is a microcosm of all cause/effect relations that can be demonstrated by deductive reasoning or numerical simulation.

In a computerized version of Occam's Razor, the hypotheses are replaced by alternative programs for a universal computer to compute a particular digital (or digitized) object X.



The shortest program is most plausible, so its *run time* measures the object's logical depth, or plausible amount of computational work required to create the object.

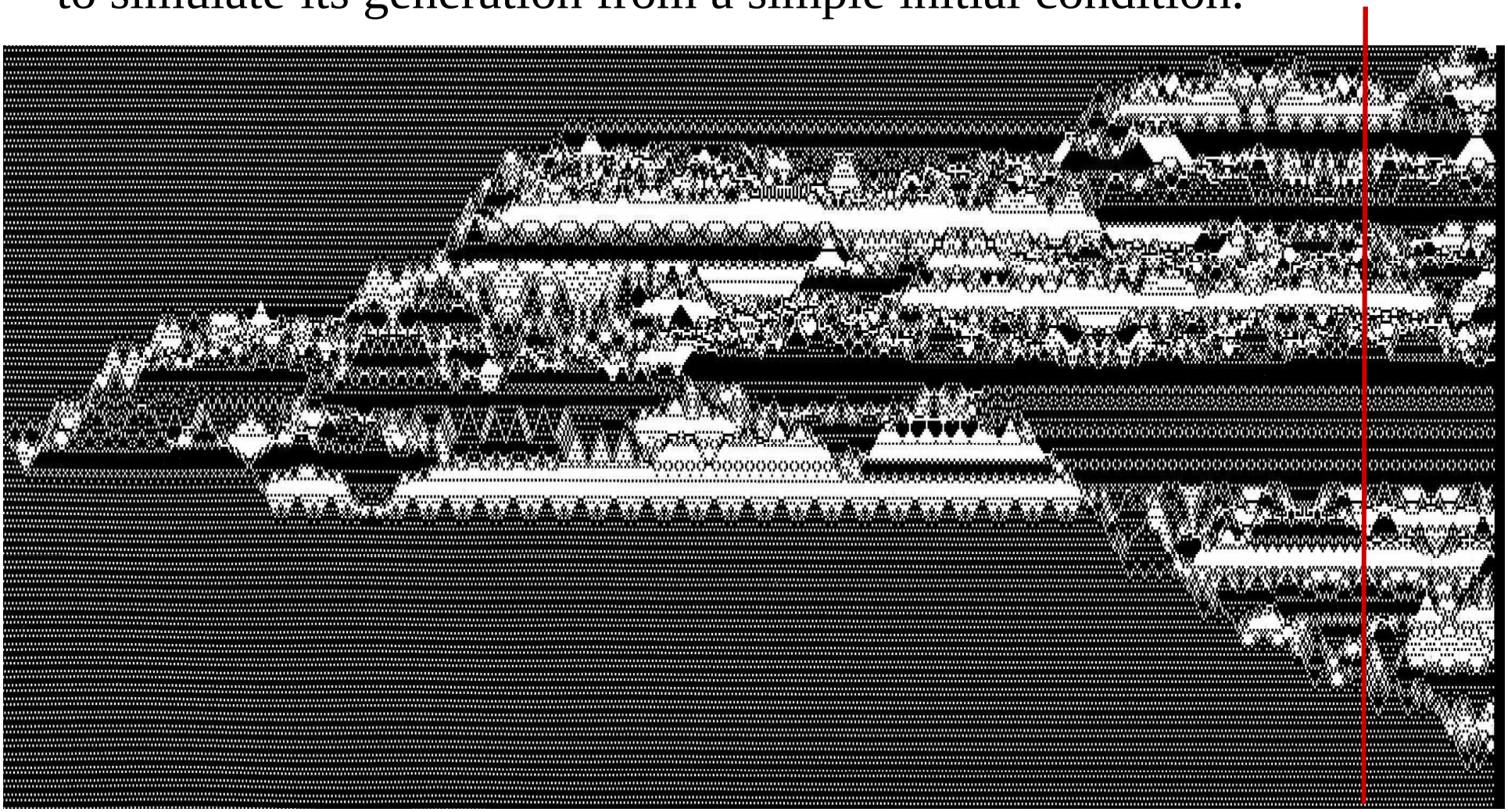
A trivially orderly sequence like 111111… is logically shallow because it can be computed rapidly from a short description.

A typical random sequence, produced by coin tossing, is also logically shallow, because it essentially **its own** shortest description, and is rapidly computable from that.

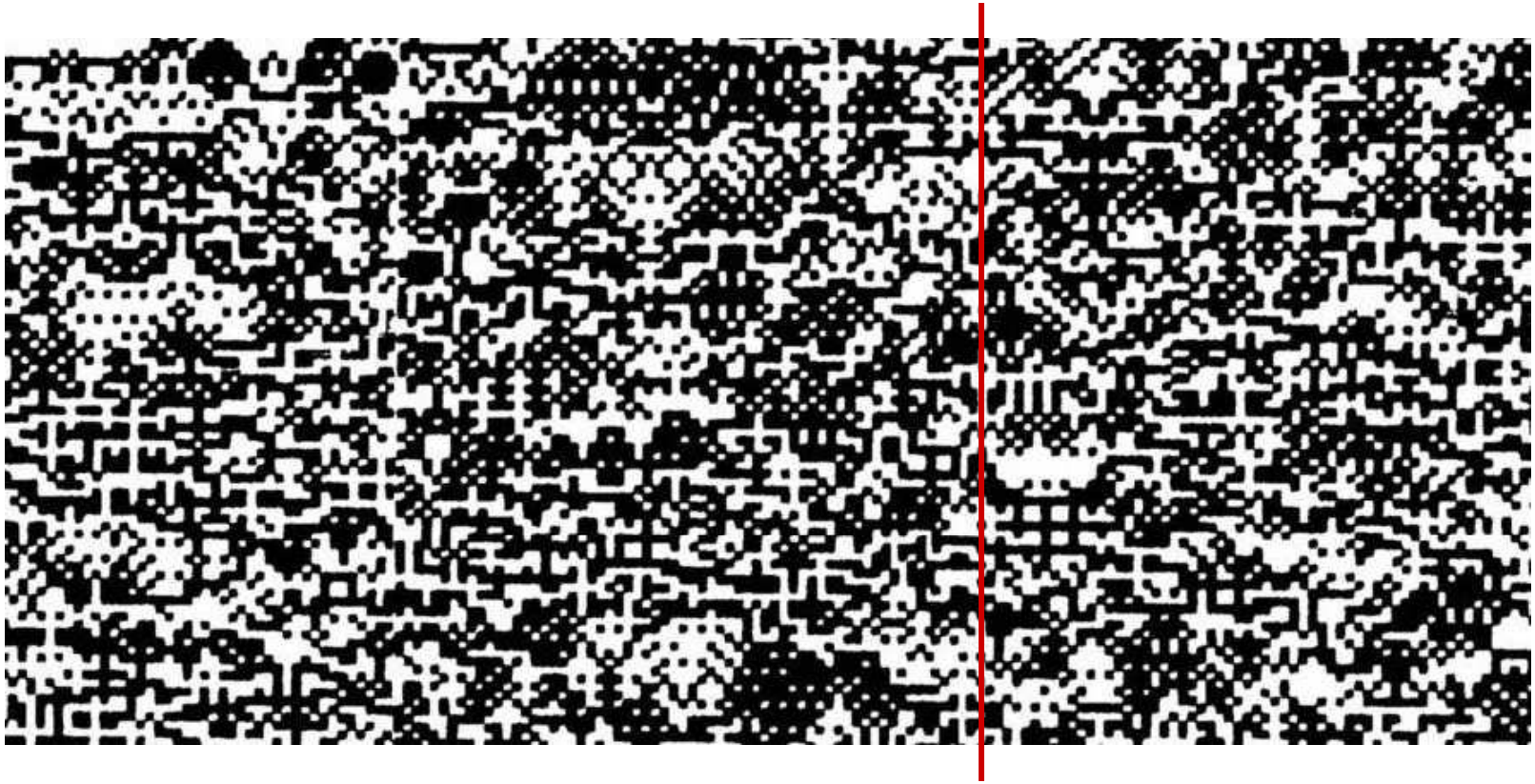
Trivial semi-orderly sequences, such as an alternating sequence of 0's and random bits, are also shallow, since they are rapidly computable from their random part.

(Depth is thus distinct from, and can vary independently from *Kolmogorov complexity* or *algorithmic information content*, defined as the **size** of the minimal description, which is high for random sequences. Algorithmic information measures a sequence's randomness, not its complexity in the sense intended here.)

Initially, and continuing for some time, the logical depth of a time slice increases with time, corresponding to the duration of the slice's actual history, in other words the computing time required to simulate its generation from a simple initial condition.



But if the dynamics is allowed to run for a large random time after equilibration (comparable to the system's Poincaré recurrence time, exponential in its size), the typical time slice becomes shallow and random, with only short-range correlations.



The minimal program for this time slice does not work by retracing its actual long history, but rather a short computation short-circuiting it.

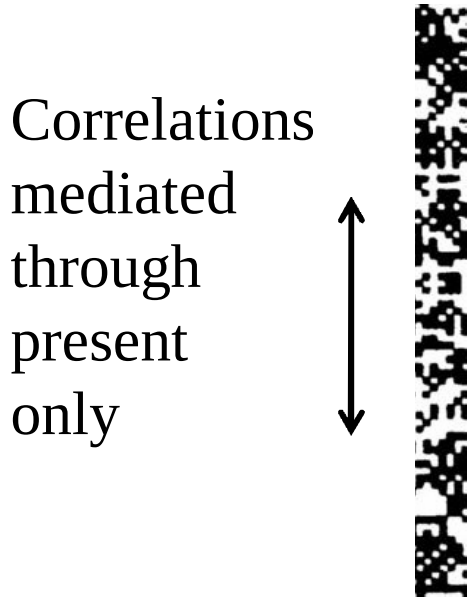
Why is the true history no longer plausible?

Because to specify the state via a simulation of its actual history would involve naming the exact **number** of steps to run the simulation.

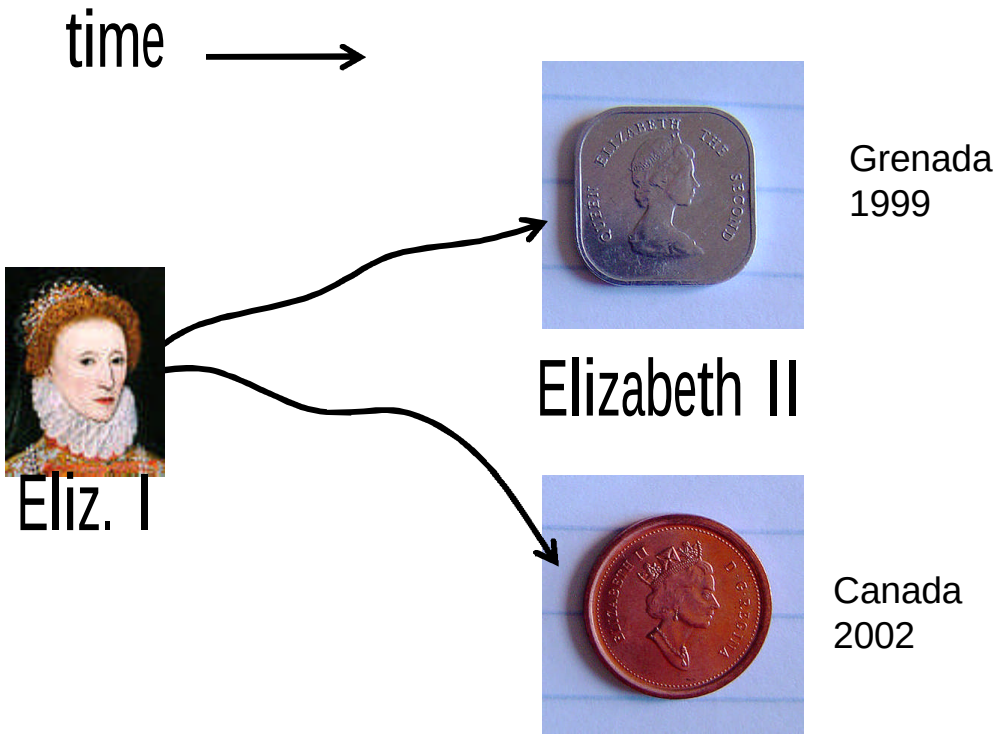
This number is typically very large, requiring about n bits to describe.

Therefore the actual history is no more plausible (in terms of Occam's razor) than a "print program" that simply outputs the state from a verbatim description.

In a world at thermal equilibrium, with local interactions correlations are generically local, mediated through the present.

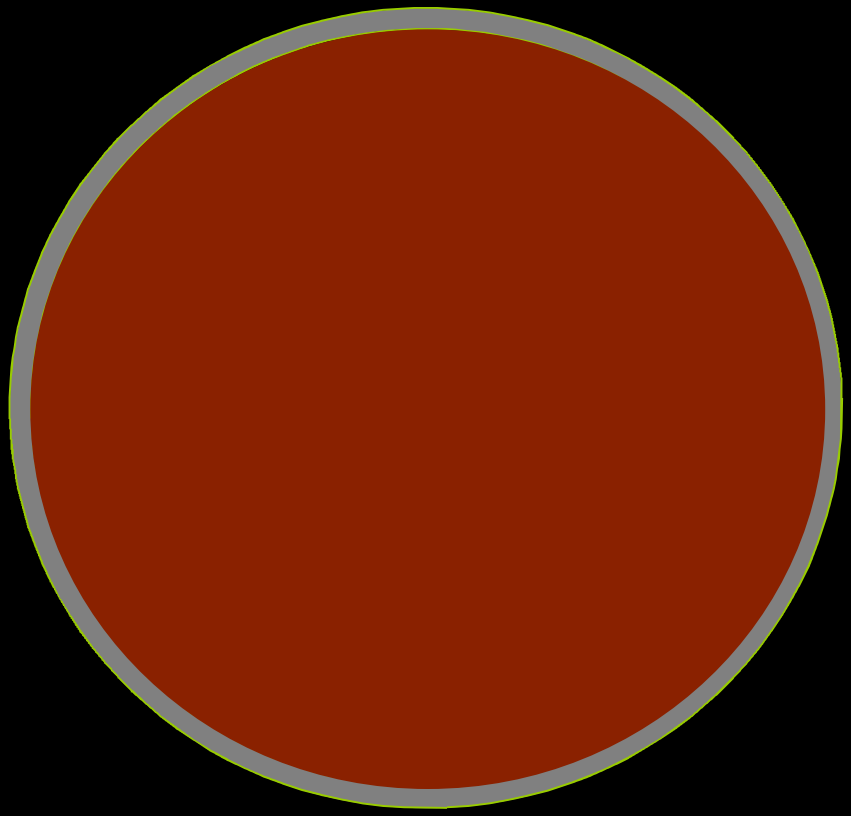


By contrast, in a non-equilibrium world, local dynamics can generically give rise to long range correlations, mediated through a V-shaped path in space-time representing a common history.



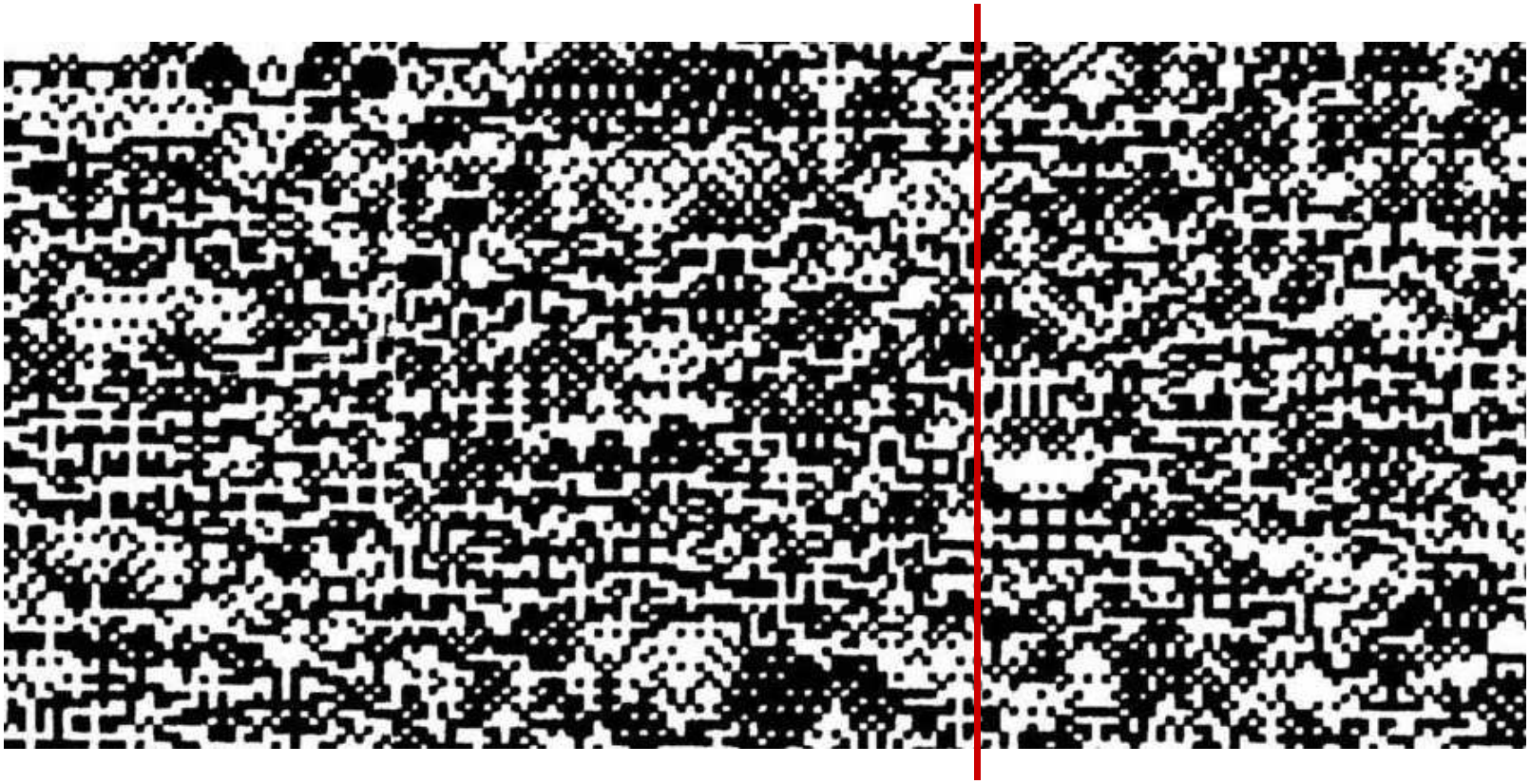
The cellular automaton is a classical toy model, but real systems with fully quantum dynamics behave similarly, losing their complexity, their long-range correlations and even their classical phenomenology as they approach equilibrium.

If the Earth were put in a large reflective box and allowed to come to equilibrium, its state would no longer be complex or even phenomenologically classical.



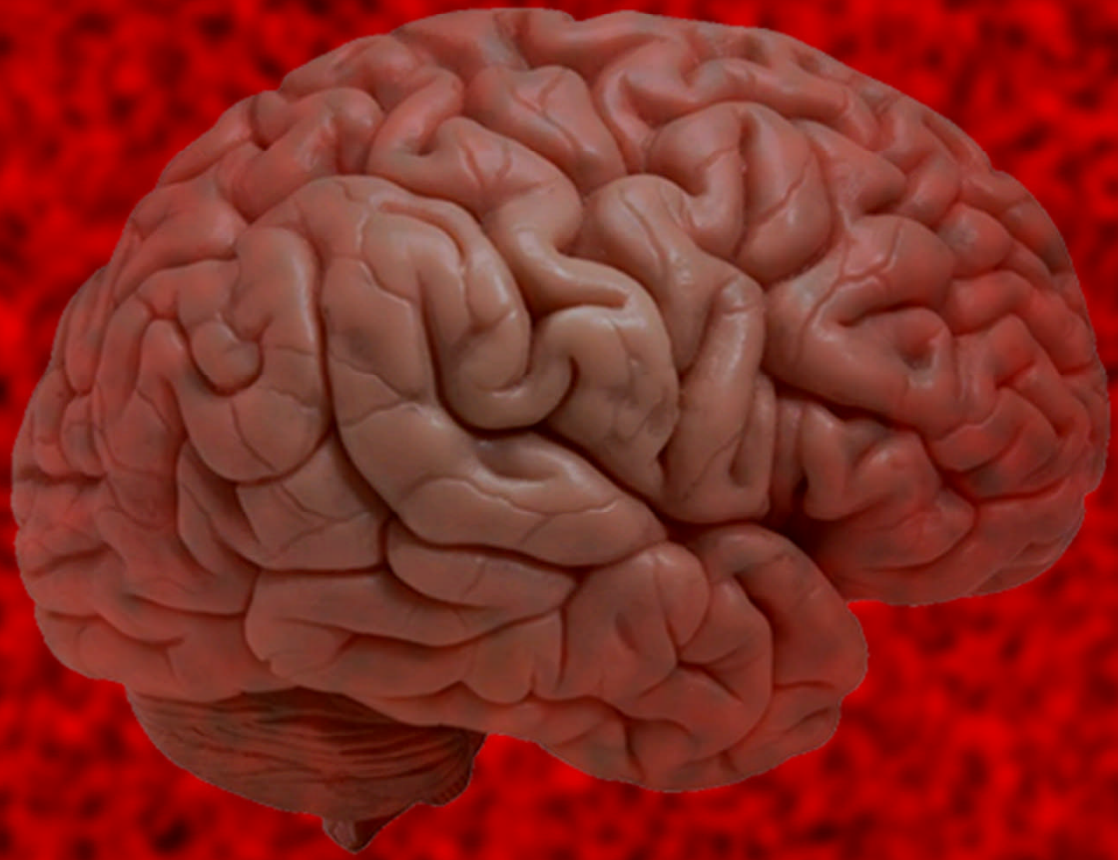
The entire state in the box would be a microcanonical superposition of near-degenerate energy eigenststates of the closed system. Such states are typically highly entangled and contain only short-range correlations.

Recall that if a system's dynamics is allowed to run for a long time after equilibration (comparable to the system's Poincaré recurrence time) its actual history can no longer be reliably inferred from its present state.



Conversely, a deep structure, one that seems to have had a long history, might just be the result of an unlikely thermal fluctuation, a so-called Boltzmann Brain.

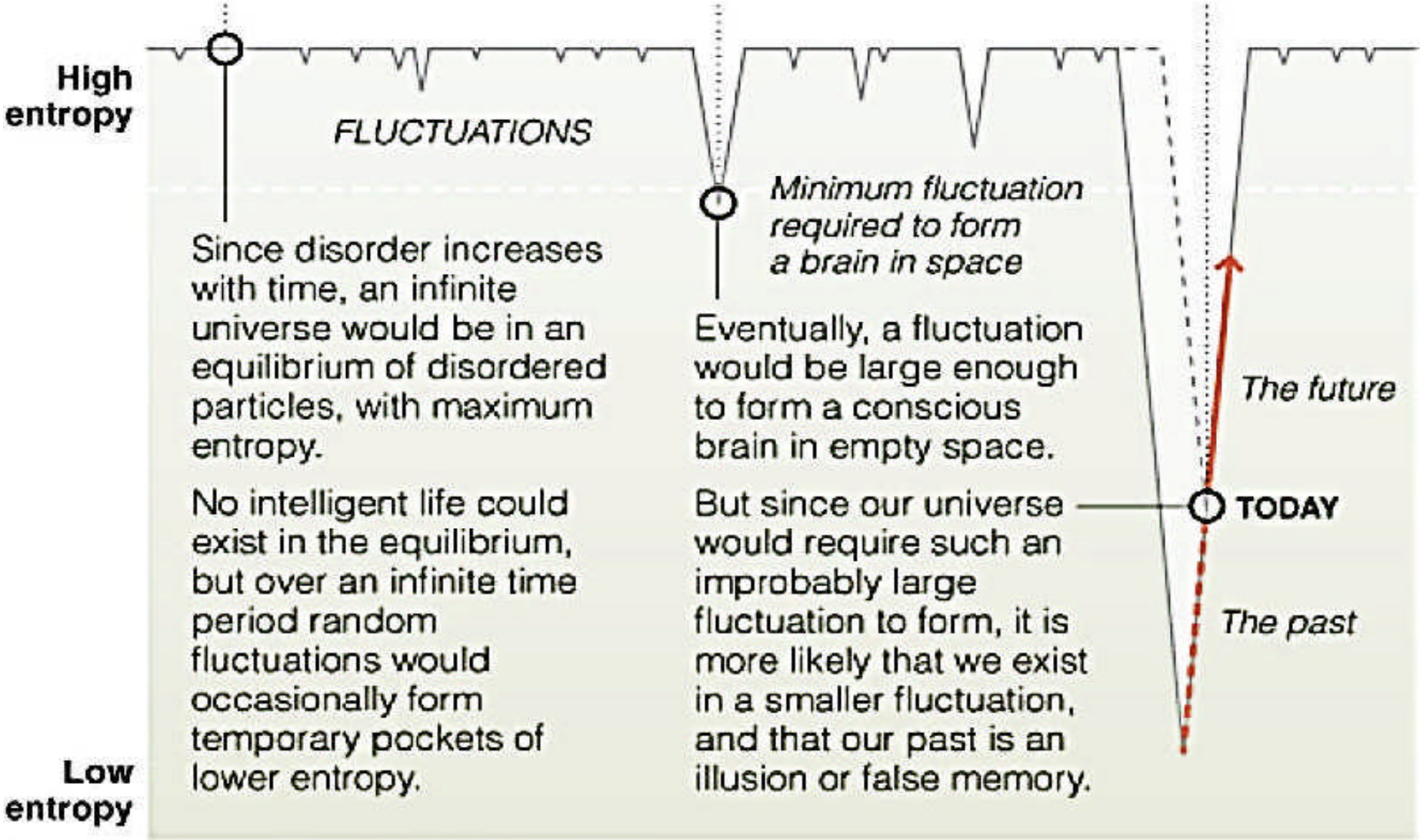
A friend of Boltzmann proposed that the low-entropy world we see may be merely a thermal fluctuation in a much larger universe. "Boltzmann Brain" has come to mean a fluctuation just large enough to produce a momentarily functioning human brain, complete with false memories of a past that didn't happen, and perceptions of an outside world that doesn't exist. Soon the BB itself will cease to exist.



Boltzmann's brain is an early example of ***anthropic*** reasoning in cosmology: arguing that what we see should be typical, not of the universe as a whole, but only of those parts of it compatible with our existence, or actually containing observers more or less like us. Of course Boltzmann imagined an infinite static universe at thermal equilibrium, whereas our cosmologies incorporate things like the Big Bang, accelerating expansion, and inflation.

It's hard to deny some validity for anthropic reasoning, but as soon as one starts using it, one gets into hard questions such as "What constitutes an observer?", and "How should observers be counted?"

A diabolical conundrum: Boltzmann fluctuations nicely explain the low entropy state of our world, and the arrow of time, but they undermine the scientific method by implying that our picture of the universe, based on observation and reason, is ***false.***



Source: Sean Carroll, California Institute of Technology

JONATHAN CORUM/THE NEW YORK TIMES

Equilibrium in a spatially or temporally infinite system , at least classically, is bad for science. We might see all sorts of amazing and beautiful things, but they'd almost certainly be private hallucinations. We would be no better off than an inhabitant of Borges' fictional Library of Babel.

On the other hand modern cosmologies offer numerous regions of long-lived disequilibrium, conducive to dissipative self-organization on a large or perhaps infinite scale.

Peter Gacs' work on dissipative fault-tolerant cellular automata, if it can be generalized to more physical environments such as field theories or eternal inflation, offers the hope of generic (on a set of positive measure in parameter space) self-organization, and even self-observing self-organization (dare we say civilization?)

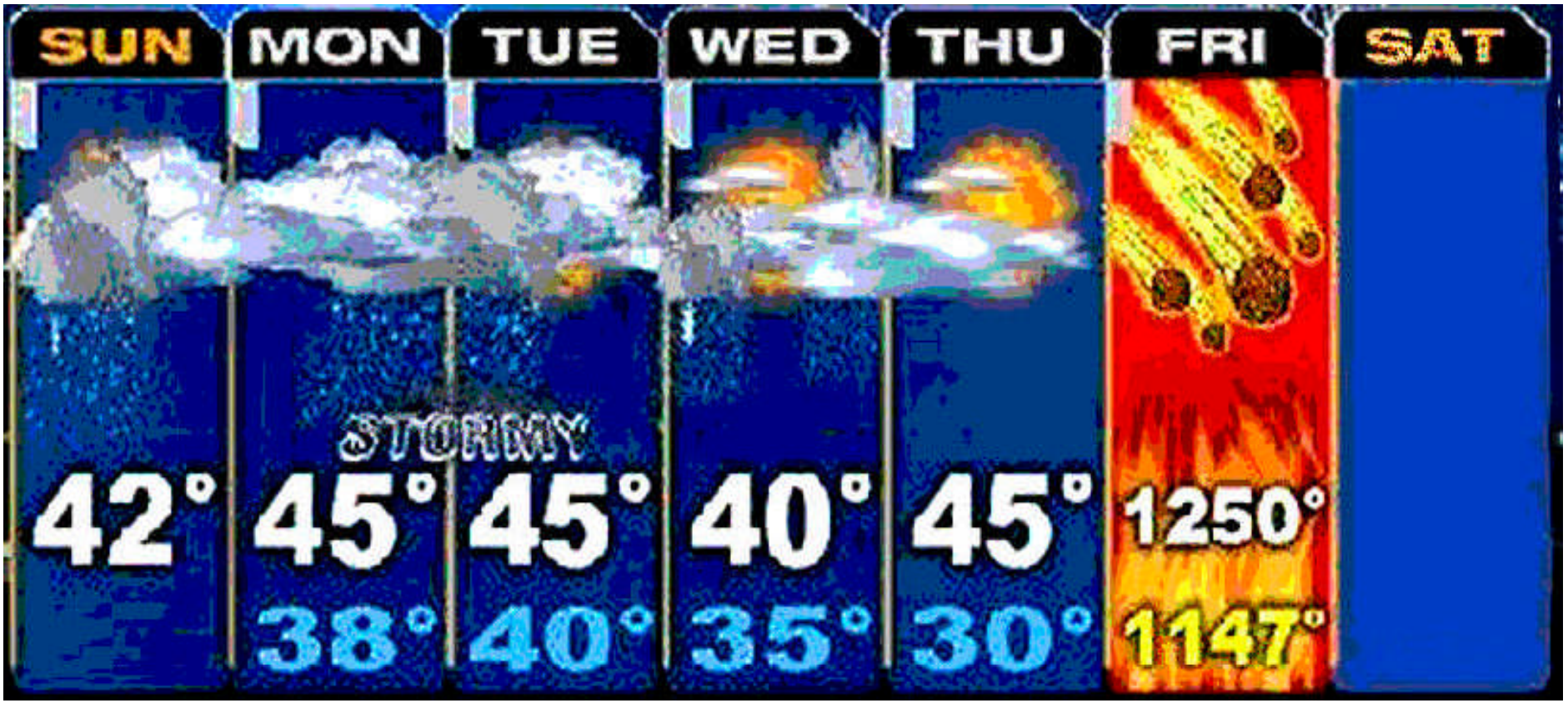
Nowadays serious cosmologists worry about Boltzmann Brains

e.g. arxiv:1308.4686

Can the Higgs Boson Save Us From the Menace of the Boltzmann Brains?

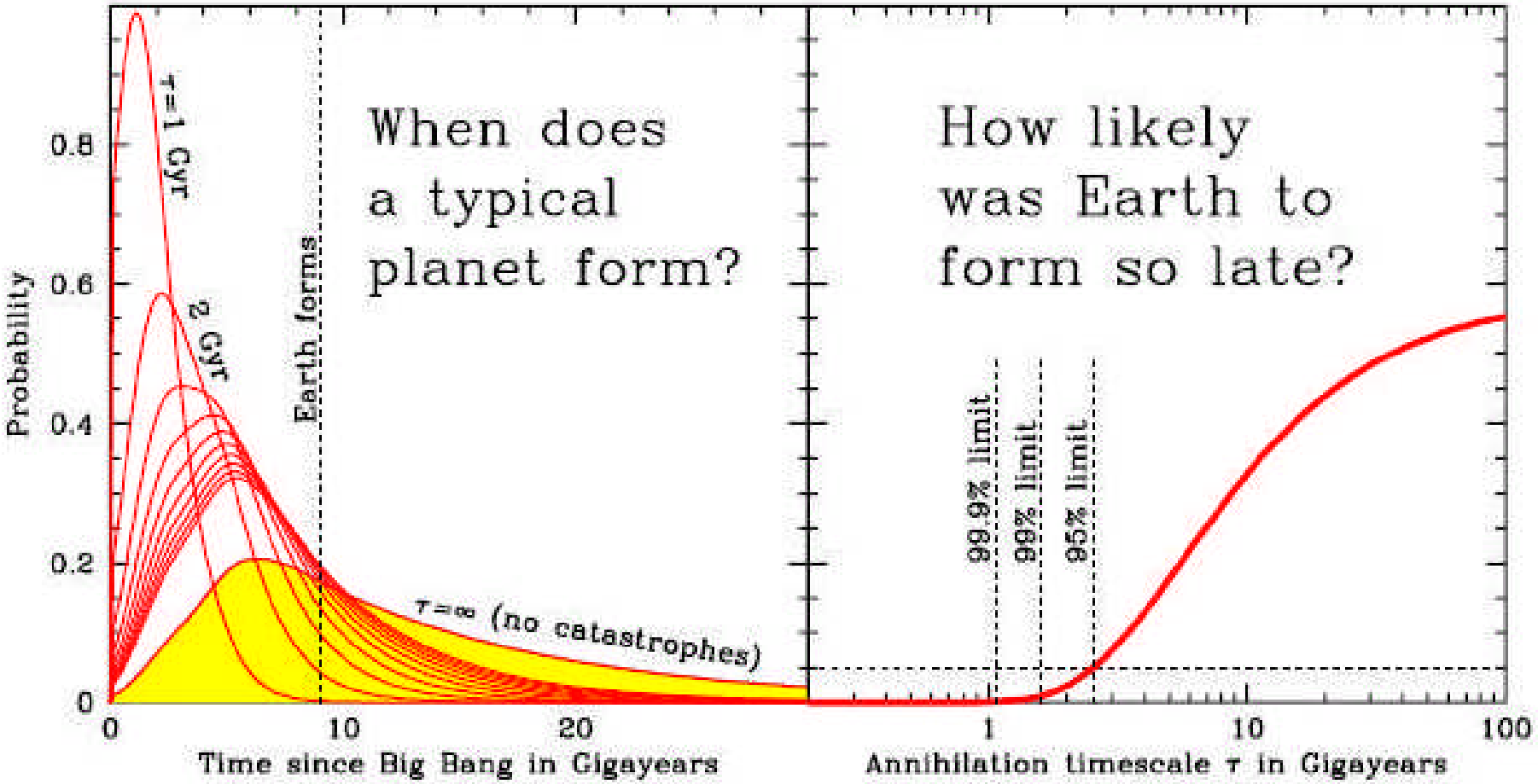
Kimberly K. Boddy and Sean M. Carroll

In other words, current cosmological models predict that the far future of our universe will be an equilibrium thermal state at positive temperature and infinite duration, giving infinitely many opportunities for Boltzmann brains to form. This seems to make it infinitely less likely that we are inhabitants of a young live universe than an old dead one. To forestall this violation of typicality, they propose that the universe will end in around 100 billion years.



Three years ago, superstitious people thought the world would end at the wraparound of the Mayan Calendar. My then 4 year old granddaughter said, “That’s silly. The world isn’t going to end.” Despite this common sense idea, it is tricky to reason about world-ending phenomena that haven’t happened yet, especially ones like Vacuum Phase Transitions that would be too sudden to notice, like dying in one’s sleep.

For example, could it be that apocalypses are intrinsically rather likely, and we've just been extraordinarily lucky so far? Tegmark and Bostrom (Nature 2005, 438, 754) argue No, on the grounds that potentially habitable planets were being formed for several billion years before the Earth.



Wigner's Friend

Schrödinger's infamous cat is in a superposition of alive and dead before the box is opened.

Eugene Wigner imagined a gentler experiment, relevant to the Quantum Boltzmann Brain problem:

Wigner's friend performs a quantum measurement with two outcomes but only tells Wigner what happened later.

After the experiment, but before Wigner hears the result, Wigner regards his friend as being in a superposition of two states, but the friend perceives only one or the other of them.

In principle (and even in practice, for atom-sized friends) Wigner can contrive for the friend to undo the measurement and forget its result—a "quantum eraser" experiment.

Wigner's friend might have been viewed as no more than a philosophical conundrum, but it is relevant to the anthropic counting of observers.

In a 2014 sequel to their 2013 paper, Boddy and Carroll, joined by Pollack, argue that it is not necessary for the universe to self-destruct to avoid the menace of Boltzmann brains. They instead argue that the late thermal state of the universe doesn't generate any Boltzmann brains because there is no mechanism to observe them, in the strong sense of making a permanent external classical record.

But as I have argued, all our experience, like that of Wigner's friend, is potentially impermanent. Therefore I think it is unreasonable to insist that nothing happens until a permanent record of it is made. Moreover observership, in the anthropic sense, is an introspective property of a system, not a property of how it would behave if measured externally.

To think about this, it helps to review some basic facts about entanglement and quantum mixed states:

- A mixed state is completely characterized by its density operator ρ, which describes all that can be learned by measuring arbitrarily many specimens of the state. For an ensemble of pure states $\{p_j, \psi_j\}$, ρ is given by the weighted sum of the projectors onto these states.
- Ensembles with the same ρ are indistinguishable.
- A system **S** in a mixed state $\rho^{\mathbf{S}}$ can, without loss of generality, be regarded as a subsystem of a larger bipartite system **RS** in a pure state $\Psi^{\mathbf{RS}}$, where R denotes a non-interacting reference system.
- "Steering" Any ensemble $\{p_j, \psi_j\}$ compatible with ρ can be remotely generated by performing measurements on the R part of $\Psi^{\mathbf{RS}}$. Measurement outcome *j* occurs with probability p_j, leaving S in state ψ_j.

Jess Riedel's scenario suggesting why Boltzmann brains ought to be present in thermal states at any positive temperature, even though there is no external observer.

- Let π_{BB} be a projector onto some state representing a fluctuation, for example a copy of the Solar System pasted into a much larger patch of de Sitter vacuum.
- Any finite temperature thermal state ρ of this patch can be expressed as a weighted sum

$$\rho = \lambda\, \pi_{BB} + (1-\lambda)\, \boldsymbol{\sigma}$$

where $\boldsymbol{\sigma}$ is a thermal state "depleted" in $\boldsymbol{\pi}_{BB}$.

- An all-powerful Preparator tosses a λ-biased coin, and prepares π_{BB} or $\boldsymbol{\sigma}$ according to the outcome.
- Before departing, the Preparator takes away, in reference system **R**, a record of all this, including, for example, souvenir photos of the just-created Earth and its inhabitants.

Since this is a valid preparation of the thermal state, and keeping in mind that it is impossible in principle to distinguish different preparations of the same mixed state, it is hard to see why the inhabitants of the de Sitter patch do not have some small probability of experiencing a life resembling our own, at least for a while.

Jason Pollack's reply to this argument: our 2014 paper, alleging the absence of such fluctuations, does not apply to all thermal states, but only those purified by a reference system **R** of a particular form, so that state $\Psi^{\mathbf{RS}}$ corresponds to a de Sitter pure state of the universe.

This may be viewed as an Occam-type argument from simplicity, favoring simplicity not of the accessible system **S**, but of the inaccessible purifying system **R**.

Internal vs External views: Our suggested internal criterion for a state ρ to have nonzero participation of a Boltzmann brain state π_{BB}, namely

$$\exists \sigma, \lambda > 0: \quad \rho = \lambda\, \pi_{BB} + (1-\lambda)\, \sigma$$

is more restrictive than the usual criterion that ρ have a positive expectation when subjected to an external measurement of π_{BB}, namely,

$$\mathrm{tr}(\rho\, \pi_{BB}) > 0.$$

Even a zero temperature vacuum state (e.g. Lorentz vacuum) would have a positive Boltzmann brain probability when measured externally. The energy for creating the Boltzmann brain out of the ground state would come from the measuring apparatus.

Cosmologists worry about typicality, especially in connection with eternal inflation, where it is hard to find a non-pathological prior distribution over "all possible universes"

D. Page, *Typicality Defended* hep-th arxiv:707.4169
• A. Garriga and J. Valenkin *Prediction and Explanation in the Multiverse* hep-th arxiv:0711.2559v3

Cosmological models like eternal inflation resemble the rest of science in being based on evidence acquired from observation and experiment.

But could one instead try to define the set of "all possible universes" in a purely mathematical way, untainted by physics?

Yes– use the universal probability defined by the Monkey Tree, despite its being only semicomputable.

(cf Juergen Schmidhuber *Algorithmic Theories of Everything* arXiv:quant-ph/0011122)

But before going so far, do we want to include any "universal" *physical* principles in the universal prior?

- Reversibility?
- Superposition – quantum mechanics
- Locality / field theories? (cf S. Lloyd and O. Dryer *The Universal Path Integral,* arxiv:1302.2850)
- Fault tolerance ≈ lack of need for fine tuning of parameters

Having made these decisions, and thus banished biology and even most of physics from the prior, what do we use as a non-anthropocentric criterion of observership?

- *Computational Universality?* Too easy: the monkey tree is already computationally universal. Similarly for unlimited depth-production

- *Science?* Formalized perhaps as Gell-Mann's IGUSes (information-gathering and utilizing systems), for example the existence of a structure that contains a more or less detailed map or explanation of its environment, and evidence of its progressive improvement.

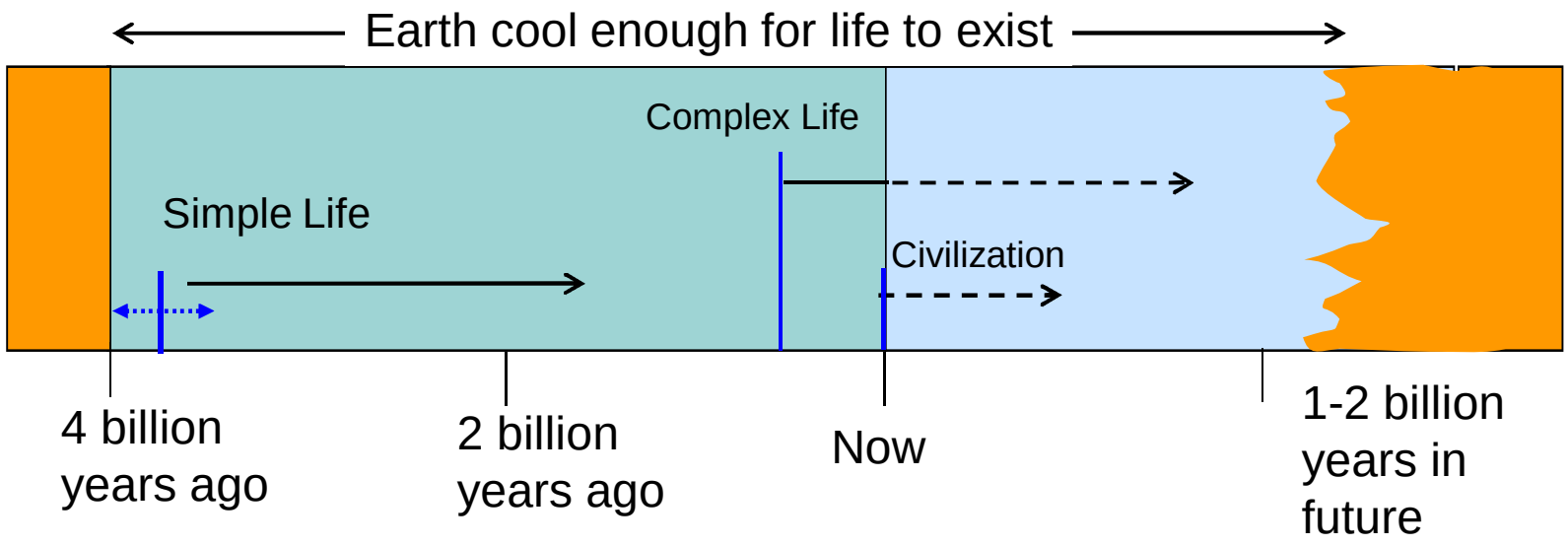
Doomsday arguments illustrate undisciplined thinking based on assumed typicality of the observer, without considering ways in which the observer may be atypical.

> "I am typical; therefore it is probable that between 5 and 95 per cent of all people who will ever live already have."

Carlton Caves' birthday party rebuttal the doomsday argument arxiv:0806.3538: Imagine wandering into a birthday party and learning that the celebrant is 50 years old. Then there is a 1/2 chance she will live to be 100 years old and a 1/3 chance to 150. Conversely, upon encountering a one day old baby, would it be fair to warn the parents that their child will probably only live a few weeks?

In both cases the person's body contains internal evidence of their life expectancy that invalidates the assumption of typicality.

A more severe doomsday question occurs in connection with *civilization*, which has existed only a *millionth* of the time potentially available for it (e.g. before the sun gets too hot).



Why is civilization so atypically new?

- VPTs? No. By Tegmark and Bostrom's argument, VPTs don't happen often enough to explain such extreme newness.
- Intrinsic Instability? Maybe civilization, especially technological civilization, may be unstable, tending to destroy itself within a few thousand years.
    - Why can't we protect ourselves from this, eg by becoming more peaceful and cooperative, or colonizing space?
    - Why don't we see the remains of previous civilizations? Maybe they're too rare, less than 1 per galaxy, which would also explain Fermi's paradox (the lack of contact with extraterrestrials).
- Perpetual newness? Maybe 1 billion years from now there will still be people, or our cultural descendants, but they will be preoccupied by some other qualitatively new feature of their existence and ask why *it* didn't happen earlier. They will still worry that by doomsday reasoning life *as they know it* may be about to disappear. (Cf. David Deutsch "The Beginning of Infinity")

In fact many people, especially dictators, fancy themselves as *atypical*, occupying a privileged temporal position at the very beginning of a long future era.

Returning to the more pessimistic hypothesis of self-destruction, Arthur Schopenhauer made an anthropic argument that we should *expect* to find ourselves in "the worst of all possible worlds." By this he meant not a world full of nastiness and evil, but one on the brink of self-destruction:

"…individual life is a ceaseless battle for existence itself; while at every step destruction threatens it. Just because this threat is so often fulfilled provision had to be made, by means of the enormous excess of the germs, that the destruction of the individuals should not involve that of the species, for which alone nature really cares. The world is therefore as bad as it possibly can be if it is to continue to be at all. Q. E. D. The fossils of the entirely different kinds of animal species which formerly inhabited the planet afford us, as a proof of our calculation, the records of worlds the continuance of which was no longer possible, and which consequently were somewhat worse than the worst of possible worlds." 1844

Open questions

- Wigner's Friend's experiences, if any
- Does entanglement enable generic fault-tolerant memory and self-organization at equilibrium (escape from Gibbs phase law)
- Are there cosmologies (e.g. eternal inflation) providing perpetual disequilibrium sufficient to support unbounded fault-tolerant classical self-organization