Title: From observers to physics via algorithmic information theory

Date: Apr 12, 2018 11:30 AM

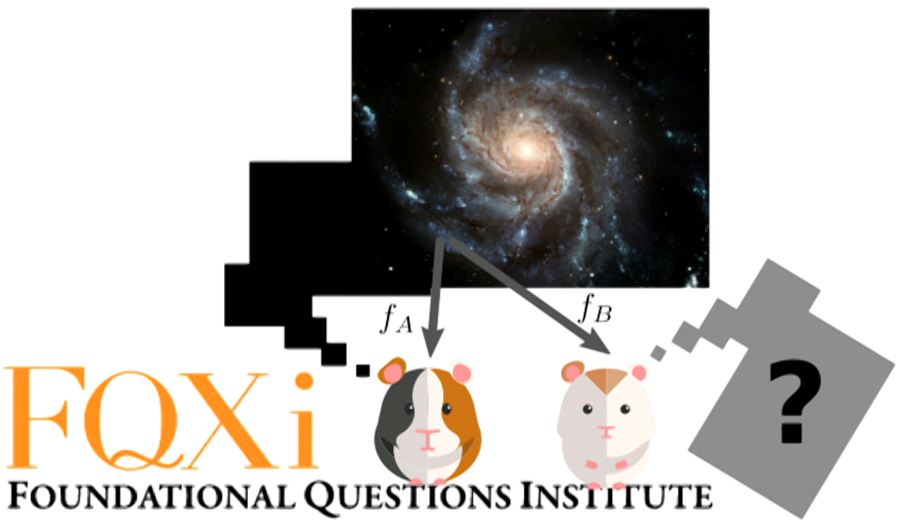
URL: http://pirsa.org/18040123

Abstract: Remark to last week's participants: This will be a condensed version of last week's talks. I will drop many details (in particular on the relation to quantum theory) and also drop the introductory slides to algorithmic probability (for this, see Marcus Hutter's introductory talk on Tuesday afternoon, April 10).

Motivated by the conceptual puzzles of quantum theory and related areas of physics, I describe a rigorous and minimal â€œproof of principleâ€• theory in which observers are fundamental and in which the physical world is a (provably) emergent phenomenon. This is a reversal of the standard view, which holds that physical theories ought to describe the objective evolution of a unique external world, with observers or agents as derived concepts that play no fundamental role whatsoever. Using insights from algorithmic information theory (AIT), I show that this approach admits to address several foundational puzzles that are difficult to address via standard approaches. This includes the measurement and Boltzmann brain problems, and problems related to the computer simulation of observers. Without assuming the existence of an external world from the outset, the resulting theory actually predicts that there is one as a consequence of AIT â€” in particular, a world with simple, computable, probabilistic laws on which different observers typically (but not always) agree. This approach represents a consistent but highly unfamiliar picture of the world, leading to a new perspective from which to approach some questions in the foundations of physics.

From observers to physics
via algorithmic information theory

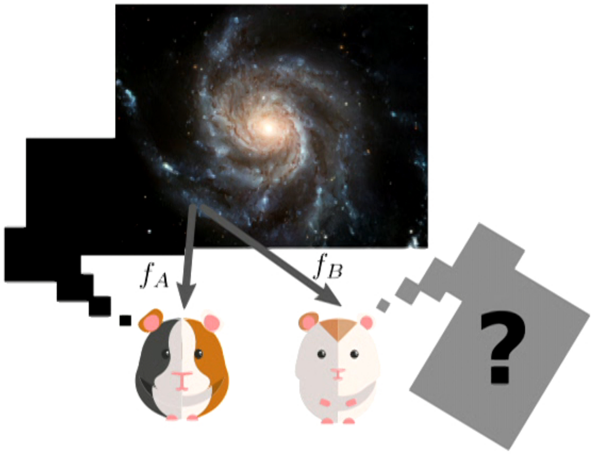
Markus P. Müller
Institute for Quantum Optics and Quantum Information, Vienna
Perimeter Institute for Theoretical Physics, Waterloo



Philosophy aspects
with Mike Cuffaro

Outline

1. Motivation



2. Postulates of the theory

3. How does an external world emerge?

4. What about more than one observer?

Systematic conceptual problems

- **Quantum theory:** measurement problem, Bell's Theorem, "no-go results for facts of the world"
- **Cosmology:** probabilities in a "big" universe (Boltzmann brains), why low-entropic initial conditions, measure problem

- **Future technology:** computer simulation of observers, simulation hypothesis, …
- **Philosophy:** Hume's problem of induction, Goodman's "new riddle", hard problems in the Philosophy of Mind

- **Naive human curiosity:** why is there a "world" with (simple, probabilistic, computable) "laws" in the first place?

Claim: These all point in a particular direction: an approach where not a "world", but **observers/observations are fundamental**.

Fundamental: **P**(*future observations* | *past observations*).

Boltzmann brain problem

Cosmologists argue about this:

"Wow! I hope I'm not, like, a disembodied brain randomly formed complete with false memories of an existence I never really had, floating in a sea of chaos and disorder. That would really ruin my day...

https://wallacegsmith.wordpress.com/2013/06/10/invasion-of-the-boltzmann-brains/

Sketch of argumentation:

- Fix a cosmological model **X** that predicts a *very* large universe.
- Count N_{BB} (# of Boltzmann brains) and compare to N_{nat} (# of naturally evolved brains).
- If $N_{BB} \gg N_{nat}$ then a "BB-observation" should be highly probable: *"What the…? I'm in space?! Aargh…"*
- That's not what we see, hence **X** is falsified.

Is this argumentation valid?
→ **what probability** should you assign to a "BB-observation"?

1. Motivation

General approach

Approach:

- Drop any assumption of an "external world".

General approach

Approach:

- Drop any assumption of an "external world".
- Start with the first-person conditional probabilities

  $$\mathbf{P}(\textit{next state of observer} \mid \textit{previous states of observer}),$$

  privately for every single observer.
- Postulate **P=algorithmic probability**, motivated by structural arguments. See what follows, and compare with actual physics.

Disclaimer

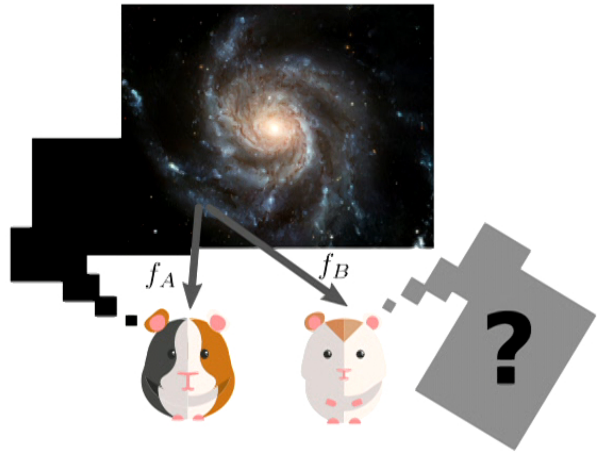
- "Observer" is a technical / information-theoretic notion. Not (directly) related to "consciousness" etc.
- Not meant as a "TOE". Predicts its own limitations. Useless for most questions that physicists care about.
- "Reality" of world is not denied, but only its fundamentality. Reproduces standard view to good approximation.

Blueprint / **proof of principle** of a certain *kind* of theory

Outline

1. Motivation

2. Postulates of the theory



3. How does an external world emerge?

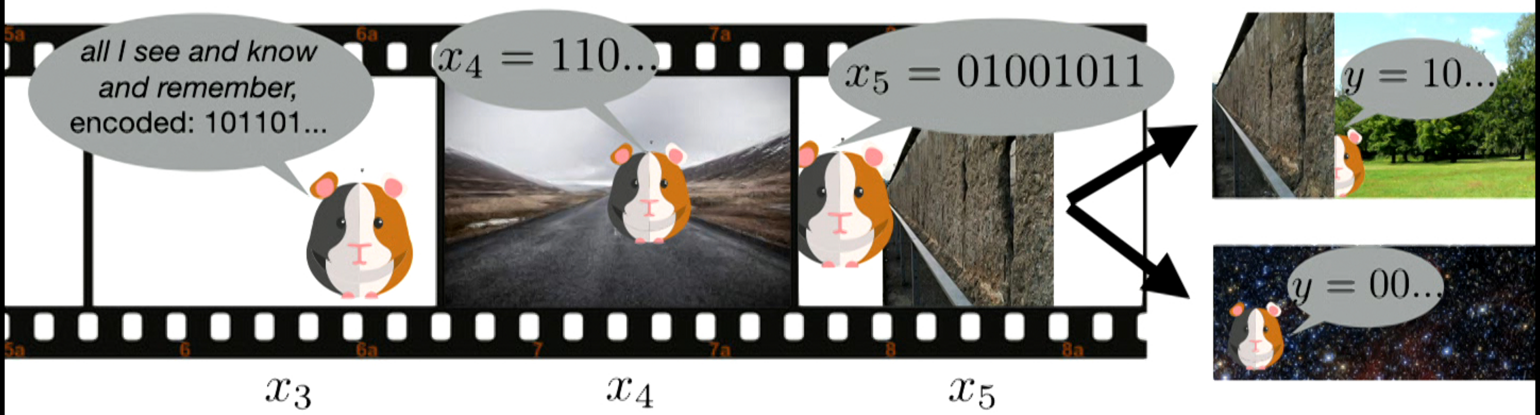
4. What about more than one observer?

Postulates of the theory

Absolutely minimal ingredients:

- An observer is in some state *x* (at any given moment).
- It will be in some other state *y* next.
- Some future states *y* are more probable than others.

→ stochastic process.

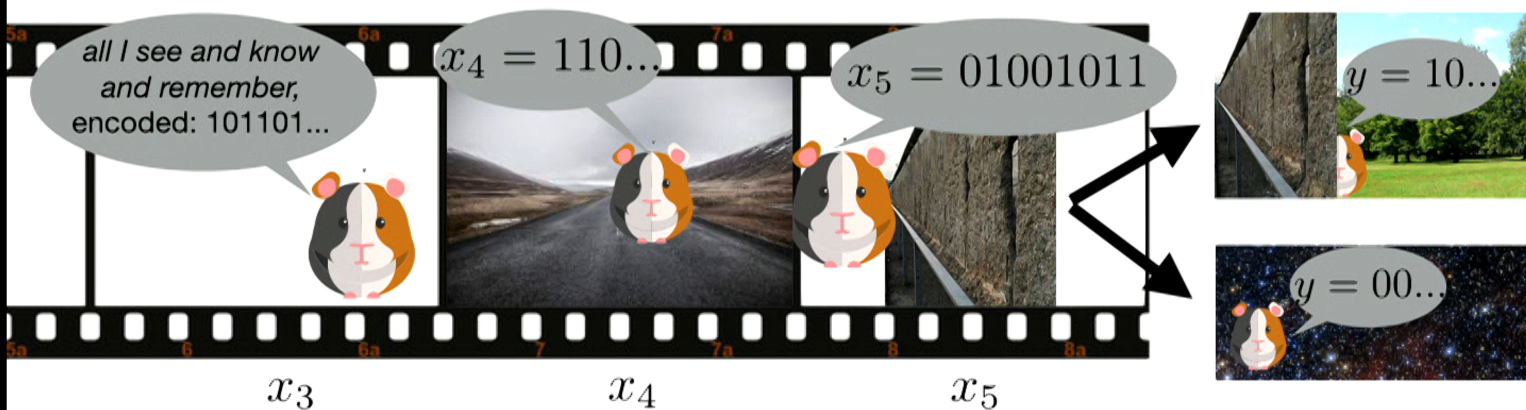


Postulates of the theory

An observer's **state** can be represented by a binary string (like $x_1 = 011010$). One (subjective) moment after the other, this yields a sequence $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, and the probability of the next state *y* is

$$\mathbf{P}(y|x_1, x_2, \ldots, x_n),$$

where **P** is conditional **algorithmic probability**.



Postulates of the theory

An observer's **state** can be represented by a binary string (like $x_1 = 011010$). One (subjective) moment after the other, this yields a sequence $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, and the probability of the next state *y* is

$$\mathbf{P}(y|x_1, x_2, \ldots, x_n),$$

where **P** is conditional **algorithmic probability**.

- No assumption that this comes from incomplete knowledge / quantum state /… of any "external world".
  The **P** describes fundamental irreducible chances.
- Not the actual 0-1-sequence is relevant, but the **computability structure** that relates the different strings. **Analogy:** in GR, the actual coordinates don't matter, but the differentiable structure.

What is algorithmic probability?

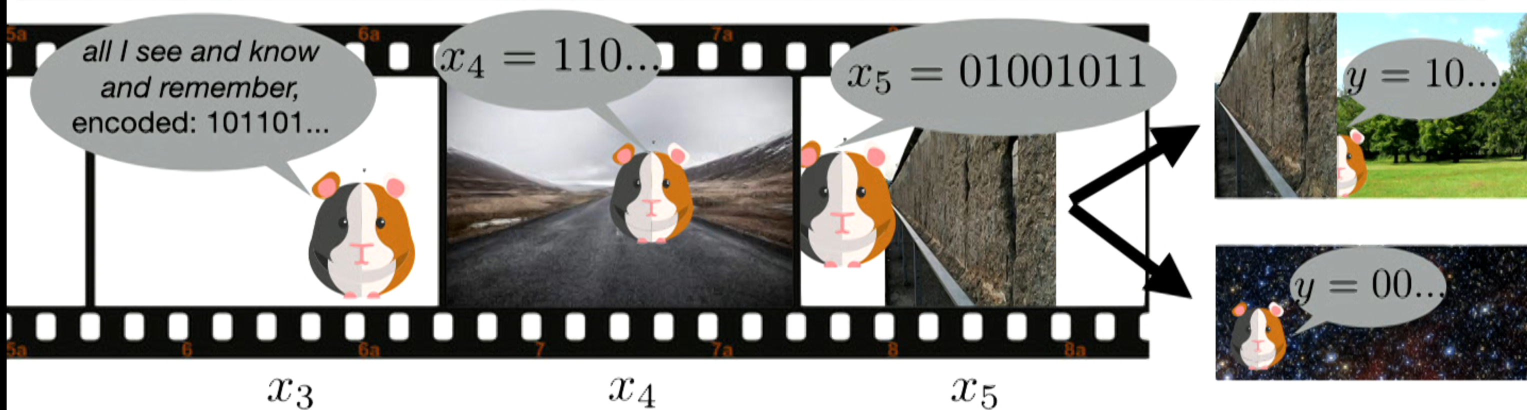
Probability measures on “histories”: $\mu(x_1,\ldots,x_n)=?$

(Boring) example: $\mu(x_1):=2^{-2\ell(x_1)-1}$, e.g. $\mu(1011)=2^{-2\cdot 4-1}=2^{-9}$,

$$\mu(x_1,\ldots,x_n):=\mu(x_1)\cdot\mu(x_2)\cdot\ldots\cdot\mu(x_n).$$

Measure: $\sum_{x_1}\mu(x_1)=1,\quad \sum_{x_{n+1}}\mu(x_1,\ldots,x_n,x_{n+1})=\mu(x_1,\ldots,x_n).$

Semimeasure: Same with “$\leq$“ instead of “=“.



What is algorithmic probability?

Measure: $\sum_{x_1} \mu(x_1) = 1, \quad \sum_{x_{n+1}} \mu(x_1, \ldots, x_n, x_{n+1}) = \mu(x_1, \ldots, x_n).$

Semimeasure: Same with "$\leq$" instead of "=".

A (semi)measure is **computable** if there is a computer program that, on input $x_1, \ldots, x_n$ and $m \in \mathbb{N}$ outputs an $(1/m)$–approximation to $\mu(x_1, \ldots, x_n)$.

A (semi)measure is **enumerable** if there is a computer program that, on input $x_1, \ldots, x_n$ and $m \in \mathbb{N}$ outputs some approximation $\mu^{(m)}(x_1, \ldots, x_n)$ such that $\mu^{(m)} \leq \mu$ and $\lim_{m \to \infty} \mu^{(m)} = \mu$.

A **universal enumerable semimeasure M** is an enumerable semimeasure such that for every enumerable semimeasure μ there exists some constant $c > 0$ such that $\mathbf{M}(x_1, \ldots, x_n) \geq c \cdot \mu(x_1, \ldots, x_n)$.

What is algorithmic probability?

Measure: $\sum_{x_1} \mu(x_1) = 1, \quad \sum_{x_{n+1}} \mu(x_1, \ldots, x_n, x_{n+1}) = \mu(x_1, \ldots, x_n).$

Semimeasure: Same with "$\leq$" instead of "=".

A (semi)measure is **computable** if there is a computer program that, on input $x_1, \ldots, x_n$ and $m \in \mathbb{N}$ outputs an $(1/m)$–approximation to $\mu(x$…

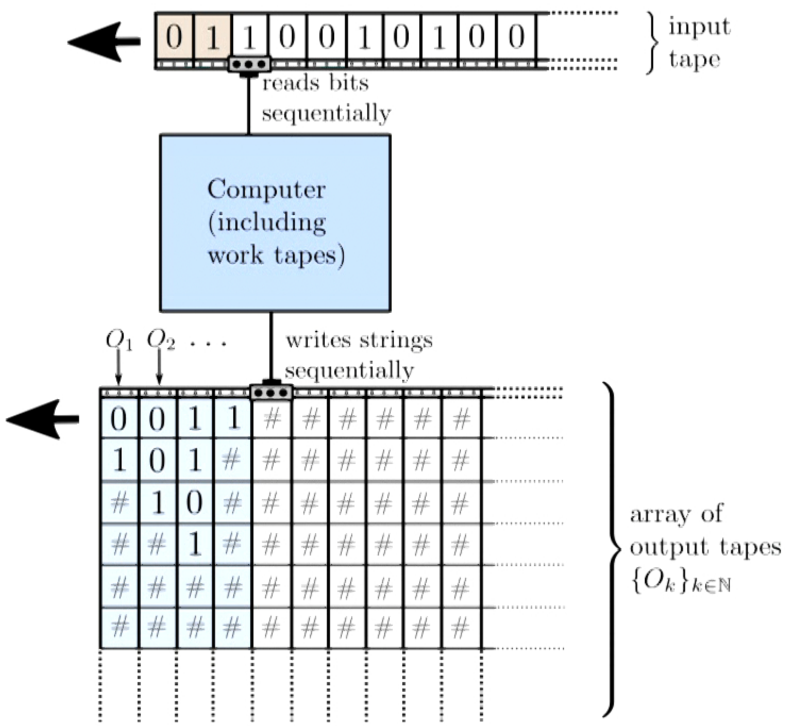
Pick any universal enumerable semimeasure **M** and normalize it.
This defines **algorithmic probability P**.

A (sem… …h that, on inp… $\mu^{(m)}(x_1, \ldots, x_n)$ such that $\mu^{(m)} \leq \mu$ and $\lim_{m\to\infty} \mu^{(m)} = \mu$.

A **universal enumerable semimeasure M** is an enumerable semimeasure such that for every enumerable semimeasure μ there exists some constant $c > 0$ such that $\mathbf{M}(x_1, \ldots, x_n) \geq c \cdot \mu(x_1, \ldots, x_n)$.

What is algorithmic probability?

Alternative definition:



Universal monotone Turing machine U

$\mathbf{M}_U$:=distribution of outputs if input is chosen at random. Is universal enumerable.

"Occam's razor":

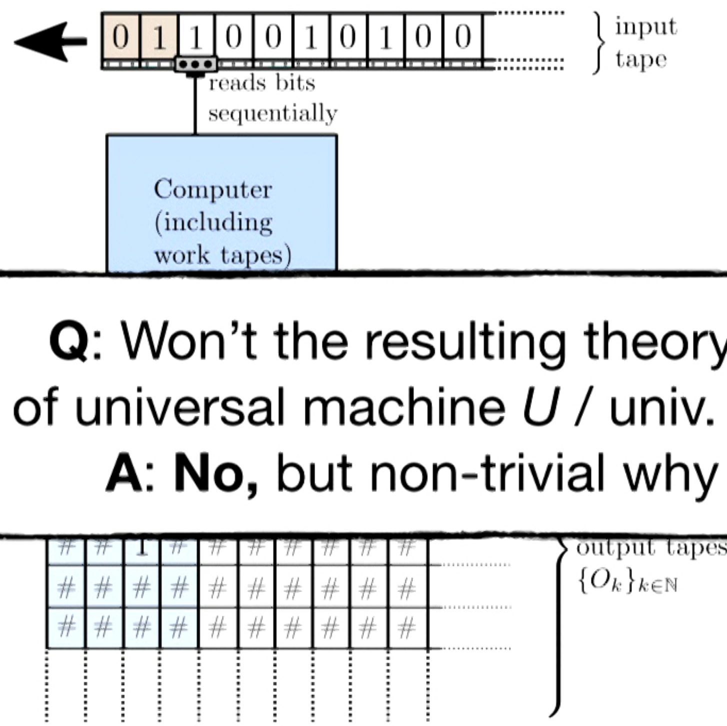
$$\mathbf{M}_U(x_1,\ldots,x_n) \geq 2^{-K(x_1,\ldots,x_n)},$$

where $K(\mathbf{x})$ is the length of the shortest computer program that outputs $\mathbf{x}$.

Favors compressibility!

What is algorithmic probability?

Alternative definition:



$\mathbf{M}_U$:=distribution of outputs if input is chosen at random. Is universal enumerable.

$x_n)$,

he

that outputs **x**.

Favors compressibility!

Q: Won't the resulting theory depend on the choice of universal machine *U* / univ. enum. semimeasure **M**?

A: **No,** but non-trivial why not. Maybe ask me later.

Universal monotone Turing machine *U*

An open problem

An observer's **state** can be represented by a binary string (like $x_1 = 011010$). One (subjective) moment after the other, this yields a sequence $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, and the probability of the next state *y* is

$$\mathbf{P}(y|x_1, \ldots, x_n) := \frac{\mathbf{P}(x_1, \ldots, x_n, y)}{\mathbf{P}(x_1, \ldots, x_n)},$$

where **P** is conditional **algorithmic probability**.

Conceptually, it would be more consequential to define **P** only to depend on the present, not the past. In some sense, the "past" is only what an observer presently remembers…

$$\mathbf{P}(y|x_n).$$

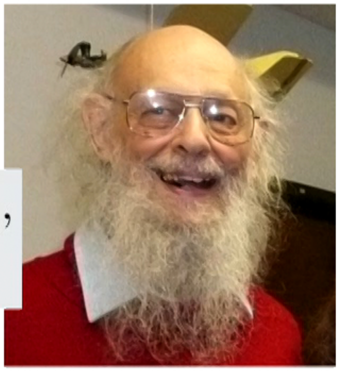
Conceptually (much) clearer, but **consequences much harder to work out**. Don't know how to do it (yet).

Why algorithmic probability?

Several possible arguments:

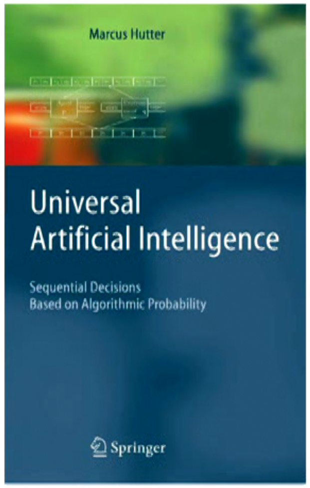
1. Extrapolating Solomonoff induction

Sol. Induction (1964): after seeing bits $b_1, \ldots, b_n$, predict the next bit b with prob. $\mathbf{P}(b|b_1 \ldots b_n)$.



Why algorithmic probability?

Several possible arguments:



1. Extrapolating Solomonoff induction

Sol. Induction (1964): after seeing bits $b_1, \ldots, b_n$, predict the next bit b with prob. $\mathbf{P}(b|b_1 \ldots b_n)$.

- Laws of physics **computable**:
  Given a description of an experiment as input, an algorithm can compute the expected outcome statistics.

- This is enough to guarantee: **Solomonoff induction will do at least as good as our best physical theories** in prediction *(in principle, asymptotically, for many observations).*

- Idea: **postulate that Solomonoff induction is "the law"!**
  This will then *have to* be consistent with physics (given our data).

Why algorithmic probability?

2. A structural motivation

Physics is nothing but what makes some future observations more likely than others.
Algorithmic probability is an essentially unique "**canonical propensity structure**".

3. A "many worlds"-like motivation

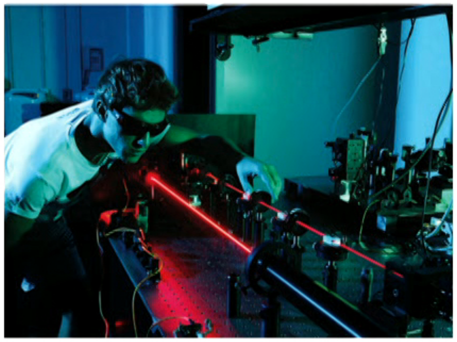
P can be interpreted as describing what an observer sees who doesn't know in which (computable) world she is located (or who is "objectively delocalized").

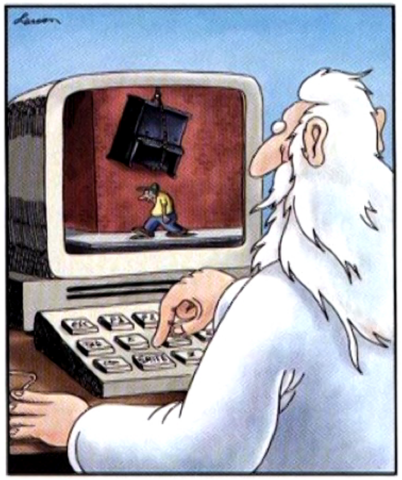
Postulate, roughly, in words

What happens to an observer is such that the best strategy for her to predict her future state would be to use the algorithmic prior — in **all** situations, no matter how crazy.

testable predictions

non-testable predictions

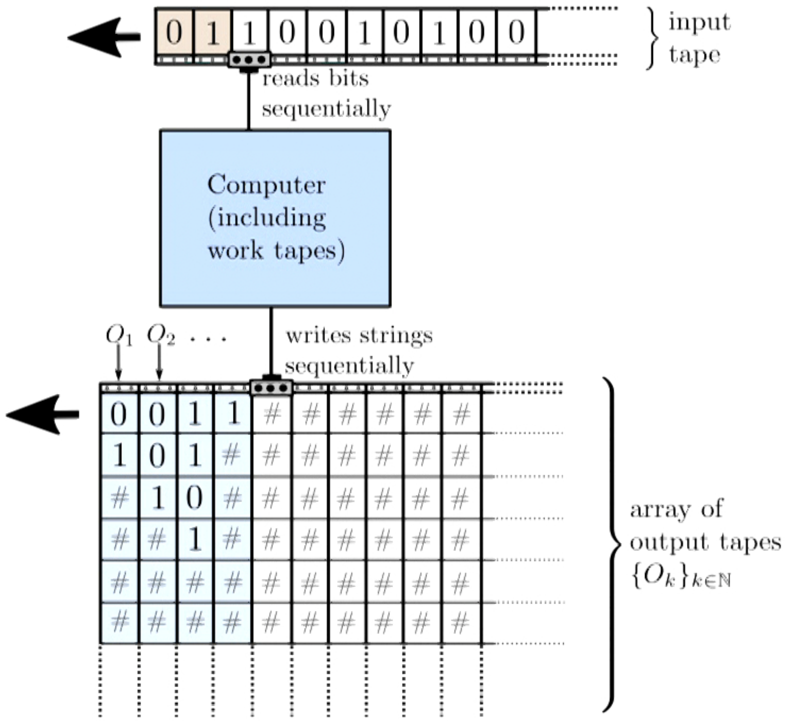




God at His computer

Like in any other physical theory, **"agency"** is **not** a fundamental part of the formalism. But can be formulated abstractly later.

Alternative definition:



Universal monotone Turing machine *U*

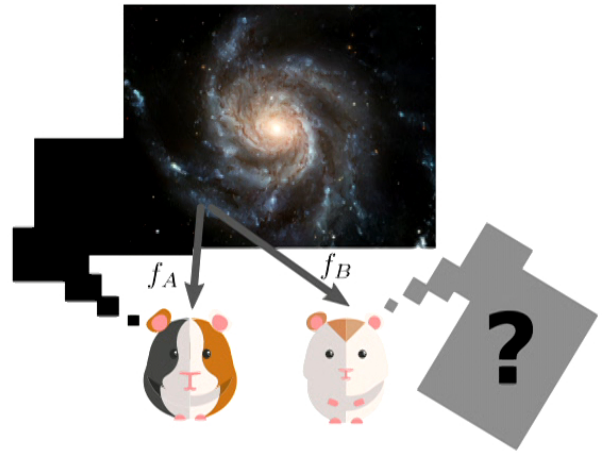
Outline

1. Motivation

2. Postulates of the theory

3. How does an external world emerge?

4. What about more than one observer?



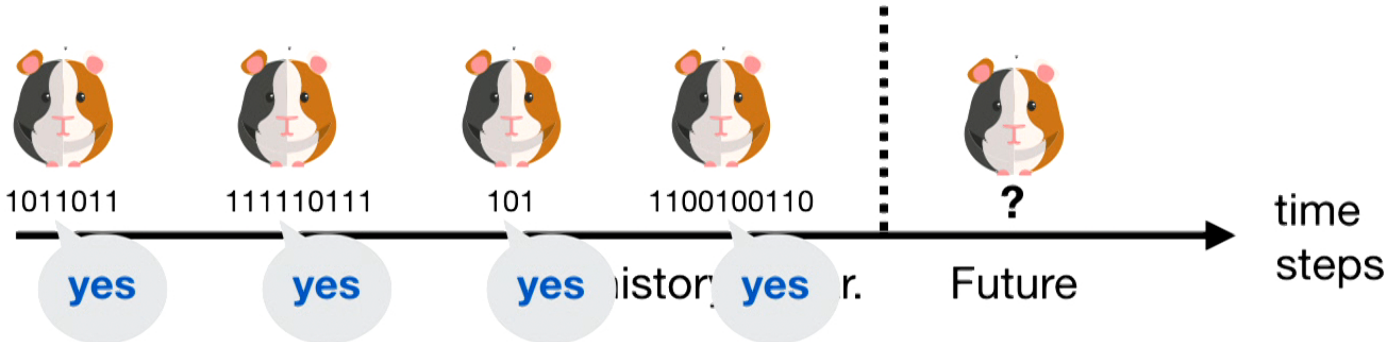
Prediction 1: **Principle of persistent regularities**

Fix any computable test *f*.

$$f(\text{bit string } x) = 0 \text{ or } 1$$

"no" "yes"

Suppose the answer has been "yes" all along:



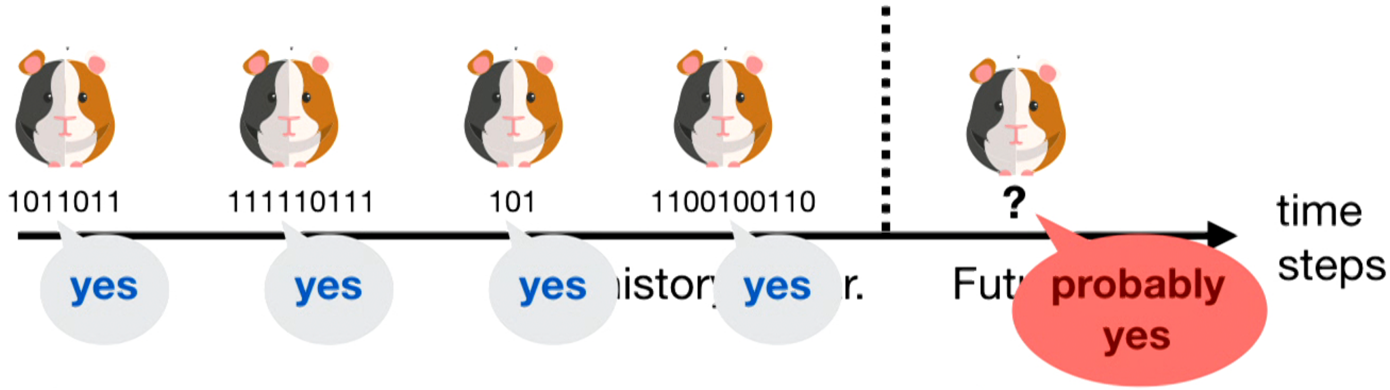
Prediction 1: **Principle of persistent regularities**

Fix any computable test *f*.

$$f(\text{bit string } x) = 0 \text{ or } 1$$

"no" "yes"

Suppose the answer has been "yes" all along:



Theorem: Then, with probability close to one, answer will be "**yes**" in the future.

Intuitive reason: This makes sequence of strings **more compressible**.

Prediction 1: **Principle of persistent regularities**

Rigorous mathematical formulation:

Theorem 8.3 (Persistence of regularities)**.** Let A be a dead-end free observer graph, and f an open computable A-test. For bits $a_1, \ldots, a_n, b \in \{0,1\}$, define the measure p as

$$p(b|a_1 a_2 \ldots a_n) := \mathbf{P}\{f(\mathbf{x}_1^{n+2}) = b \mid f(\mathbf{x}_1^2) = a_1, \ldots, f(\mathbf{x}_1^{n+1}) = a_n\},$$

and similarly define the semimeasure m with $\mathbf{P}$ replaced by $\mathbf{M}$. Then we have[38] $m(0|1^n) \leq 2^{-K(n)+\mathcal{O}(1)}$, and for the measure p we have the slightly less explicit statement

$$p(1|1^n) \stackrel{n\to\infty}{\longrightarrow} 1, \qquad (10)$$

but the convergence is rapid since $\sum_{n=0}^{\infty} p(0|1^n) < \infty$. Thus, e.g., $p(1|1^n) > 1 - \frac{1}{n}$ for all but finitely many n. Moreover, the probability that $f(\mathbf{x}_1^{n+1}) = 1$ for all $n \in \mathbb{N}$ is non-zero.

Prediction 1: **Principle of persistent regularities**

This is already indicates how **Boltzmann brains** are exorcized:

f := computable test whether observations are typical for a planet-like environment.

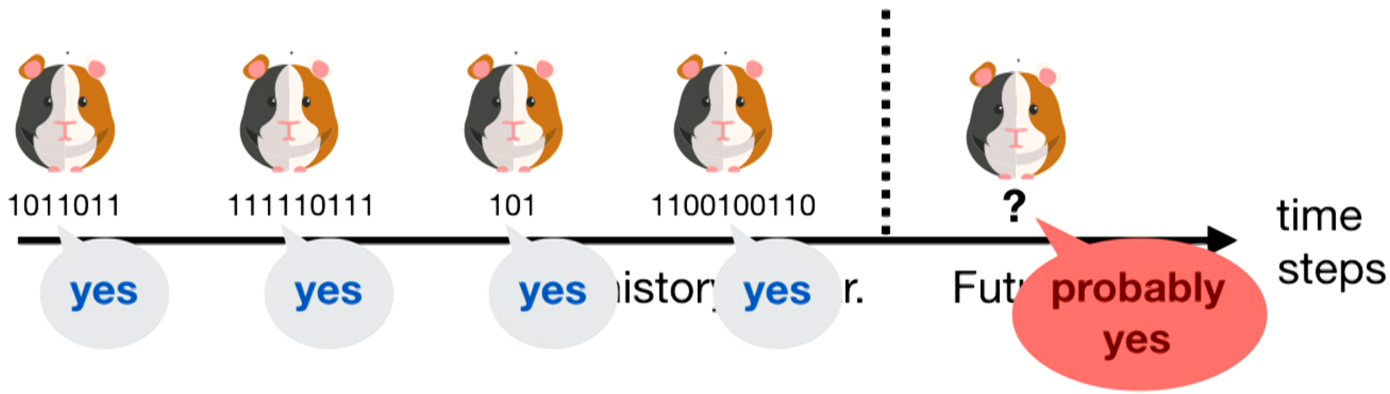
Suppose the answer has been "yes" all along:

Prediction 1: **Principle of persistent regularities**

This is already indicates how **Boltzmann brains** are exorcized:

f := computable test whether observations are typical for a planet-like environment.

Suppose the answer has been "yes" all along:



Boltzmann brain experience ("*what the... I'm suddenly in space... argh!!*") is highly unlikely.

Prediction 1: **Principle of persistent regularities**

Rigorous mathematical formulation:

Theorem 8.3 (Persistence of regularities)**.** Let A be a dead-end free observer graph, and f an open computable A-test. For bits $a_1, \ldots, a_n, b \in \{0,1\}$, define the measure p as

$$p(b|a_1 a_2 \ldots a_n) := \mathbf{P}\{f(\mathbf{x}_1^{n+2}) = b \mid f(\mathbf{x}_1^2) = a_1, \ldots, f(\mathbf{x}_1^{n+1}) = a_n\},$$

and similarly define the semimeasure m with $\mathbf{P}$ replaced by $\mathbf{M}$. Then we have[38] $m(0|1^n) \leq 2^{-K(n)+\mathcal{O}(1)}$, and for the measure p we have the slightly less explicit statement

$$p(1|1^n) \stackrel{n\to\infty}{\longrightarrow} 1, \qquad (10)$$

but the convergence is rapid since $\sum_{n=0}^{\infty} p(0|1^n) < \infty$. Thus, e.g., $p(1|1^n) > 1 - \frac{1}{n}$ for all but finitely many n. Moreover, the probability that $f(\mathbf{x}_1^{n+1}) = 1$ for all $n \in \mathbb{N}$ is non-zero.

Prediction 1: **Principle of persistent regularities**

But it is not quite enough — cf. Goodman's **New Riddle** of Induction:

f := computable test whether observations are typical for a planet-like environment.

$$\tilde f := \begin{cases} f & \text{if observed calendar shows year } \leq 2050 \\ \text{NOT } f & \text{if observed calendar shows year } > 2050. \end{cases}$$

(cf. Goodman's green/blue versus bleen/grue).

Theorem applies to both f and $\tilde f$. Contradiction?! **No.**

Resolution: Since $K(f) < K(\tilde f)$, the f-regularity stabilizes **earlier** than the $\tilde f$-regularity.

Careful quantitative analysis of *K* (see paper) confirms exorcism of the Boltzmann brains.

Prediction 1: **Principle of persistent regularities**

But it is not quite enough — cf. Goodman's **New Riddle** of Induction:

f := computable test whether observations are typical for a planet-like environment.

$$\tilde f := \begin{cases} f & \text{if observed calendar shows year } \leq 2050 \\ \text{NOT } f & \text{if observed calendar shows year } > 2050. \end{cases}$$

(cf. Goodman's green/blue versus bleen/grue).

Theorem applies to both f and $\tilde f$. Contradiction?! **No.**

Resolution: Since $K(f) < K(\tilde f)$, the f-regularity stabilizes **earlier** than the $\tilde f$-regularity.

Careful quantitative analysis of *K* (see paper) confirms exorcism of the Boltzmann brains.

Will the different regularities "fit together" coherently? **Yes!** →

Prediction 2: **Simple, computable, probabilistic "world"**

Theorem. Consider any **computable probabilistic process** that has description length *L* on a universal computer. Suppose it generates outputs $x'_1, x'_2, x'_3, \ldots$ according to the (computable) distribution $\mu(x'_1, \ldots, x'_n)$.

Then, with **P**-probability at least 2^{-L} we have

$$\mathbf{P}(y|x_1, \ldots, x_n) \stackrel{n\to\infty}{\longrightarrow} \mu(y|x_1, \ldots, x_n),$$

i.e. the outputs of this process will asymptotically be a perfect description of the observer's states.

observer state, **P**-distributed

looks as if it came from

computational process, output μ-distributed.

Prediction 2: **Simple, computable, probabilistic “world”**

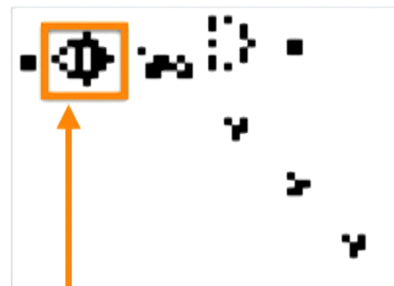
Theorem. Consider any **computable probabilistic process** that has description length L on a universal computer. Suppose it generates outputs $x'_1, x'_2, x'_3, \ldots$ according to the (computable) distribution $\mu(x'_1, \ldots, x'_n)$. Then, with **P**-probability at least 2^{-L} we have

$$\mathbf{P}(y|x_1, \ldots, x_n) \stackrel{n\to\infty}{\longrightarrow} \mu(y|x_1, \ldots, x_n),$$

i.e. the outputs of this process will asymptotically be a perfect description of the observer’s states.

observer state, **P**-distributed

looks as if it came from



computational process, output μ-distributed.

Prediction 2: **Simple, computable, probabilistic "world"**

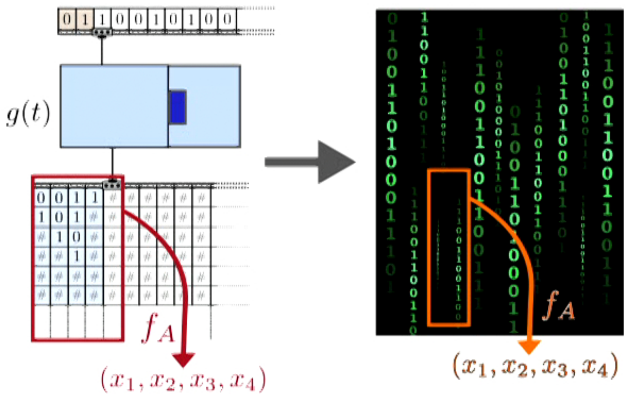
Theorem. Consider any **computable probabilistic process** that has description length L on a universal computer. Suppose it generates outputs $x'_1, x'_2, x'_3, \ldots$ according to the (computable) distribution $\mu(x'_1, \ldots, x'_n)$. Then, with **P**-probability at least 2^{-L} we have

$$\mathbf{P}(y|x_1, \ldots, x_n) \stackrel{n\to\infty}{\longrightarrow} \mu(y|x_1, \ldots, x_n),$$

i.e. the outputs of this process will asymptotically be a perfect description of the observer's states.

- It is **contingent** which process (and thus μ) will emerge, but **simpler** ones are highly preferred (simpler = smaller L = higher probability).
- Thus, observer's probabilities will equal the marginal distribution of some random variable that's part of a **probabilistic process** with **computable laws of short description** (a simple algorithm).

Prediction 2: **Simple, computable, probabilistic “world”**

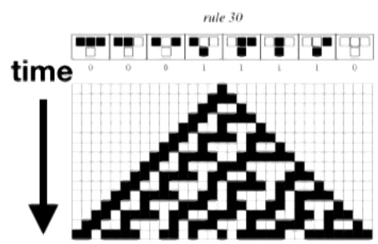


Abstract process (not even necessarily discrete in a naive sense).

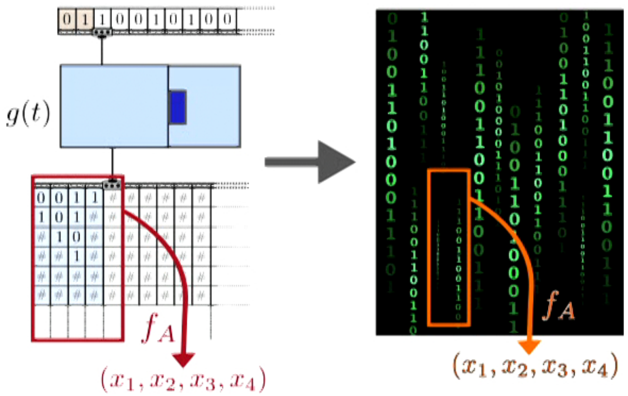
“External world”: computational ontological model, useful for predicting future experiences by providing direct causal/mechanistic explanations.

Comparison with physics that we observe:

- Generically, (simple) computations start in **simple initial state**, and then evolve with increasing algorithmic entropy.



Prediction 2: **Simple, computable, probabilistic "world"**



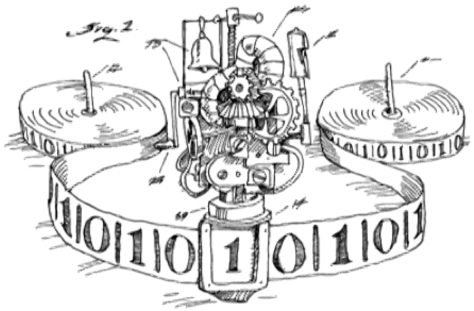
Abstract process (not even necessarily discrete in a naive sense).

"External world": computational ontological model, useful for predicting future experiences by providing direct causal/mechanistic explanations.

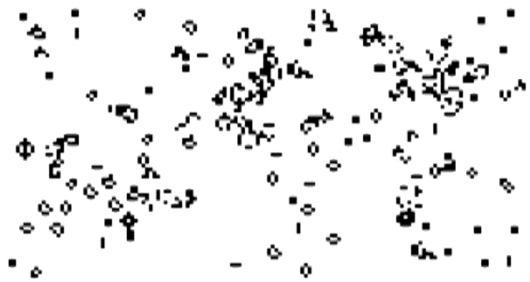
Comparison with physics that we observe:

- Generically, (simple) computations start in **simple initial state**, and then evolve with increasing algorithmic entropy.
- Time evolution is in principle **simulatable by a (short) Turing machine program** (but not necessarily efficiently!).

Prediction 2: **Simple, computable, probabilistic "world"**



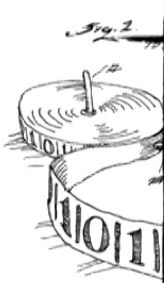
is not a very natural model of computation for



Comparison with physics that we observe:

- Generically, (simple) computations start in **simple initial state**, and then evolve with increasing algorithmic entropy.
- Time evolution is in principle **simulatable by a (short) Turing machine program** (but not necessarily efficiently!).
- Process is **fundamentally probabilistic**, but TM not necessarily the most natural model of computation to represent the process.

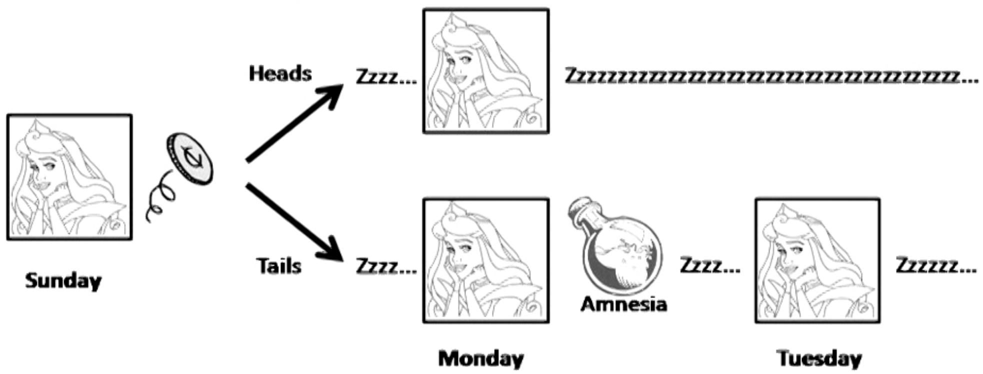
Prediction 2: **Simple, computable, probabilistic "world"**



is not

In fact, **violation of Bell inequalities and no-signalling** (wrt. computational causal structure) generic predictions, assuming that the open problem from further above can be solved.

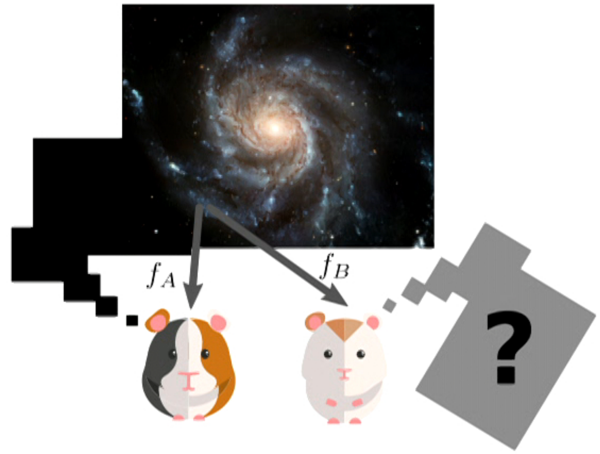
http://pirsa.org/displayFlash.php?id=18040080



Compari

- Gene … state, ✓
  and t
- Time … uring **mac** ✓
- Proc … essarily the most natural model of computation to represent the process. ✓

Outline



1. Motivation

2. Postulates of the theory

3. How does an external world emerge?

4. What about more than one observer?

Prediction 3: **An emergent notion of objectivity**

Apriori, different observers make their *own* "private" observations.

Abby

$$\mathbf{P}(y^A | x_1^A, \ldots, x_n^A)$$

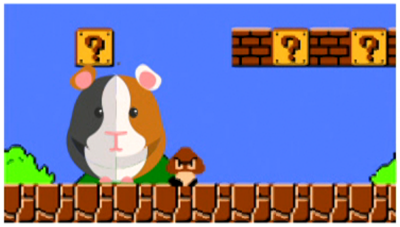
Bambi

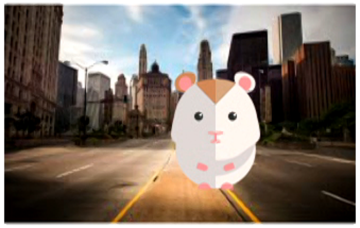
$$\mathbf{P}(y^B | x_1^B, \ldots, x_m^B)$$

Prediction 3: **An emergent notion of objectivity**

Apriori, different observers make their *own* "private" observations. They are completely unrelated, and live in their own "external worlds".

A-world





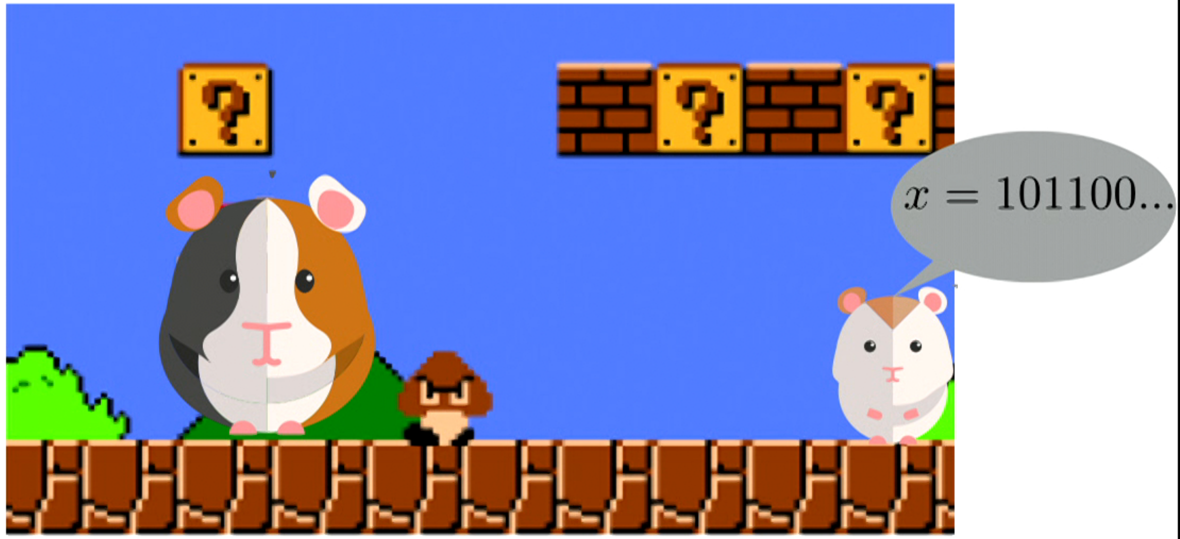
B-world

But suppose that **A** sees something in her external world that seems like another observer **B** to her...

Prediction 3: **An emergent notion of objectivity**

Apriori, different observers make their *own* "private" observations. They are completely unrelated, and live in their own "external worlds".

A-world

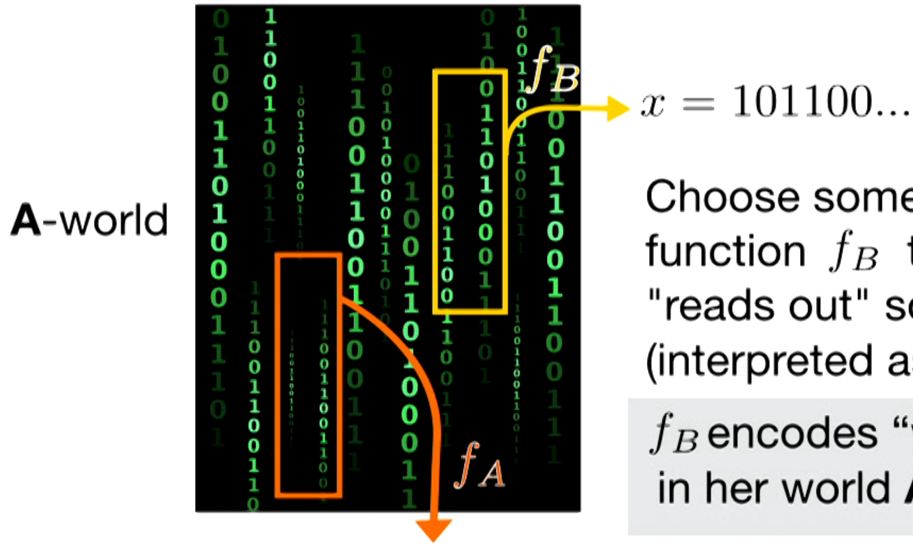


But suppose that **A** sees something in her external world that seems like another observer **B** to her...

Does what **A** sees really correspond to the first-person perspective of another observer?

Prediction 3: **An emergent notion of objectivity**

How to formalize this:



Choose some (simple) computable function f_B that, at any time step, "reads out" some binary string (interpreted as **B**'s current state)

f_B encodes "what other thing in her world **A** is looking at".

Two probability distributions:

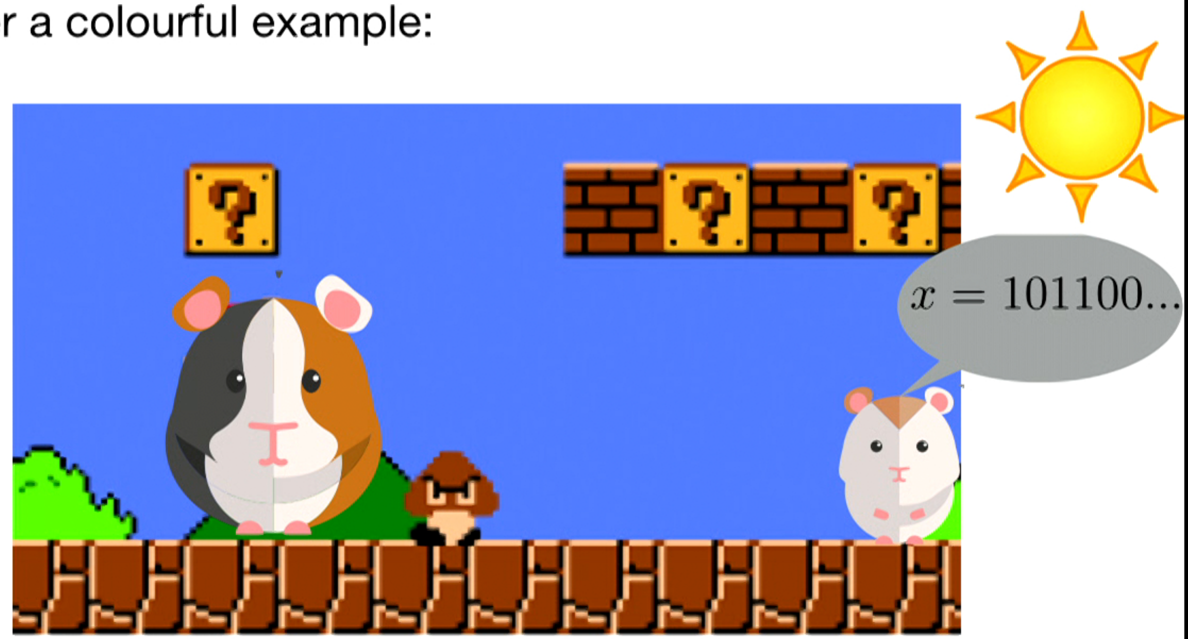
$\nu(x_1, x_2, \ldots, x_n) :=$ prob. that **B** is in states $x_1, \ldots, x_n$ acc. to **A**-world

$\mathbf{P}(x_1, \ldots, x_n) =$ algorithmic probability that **B** is in states $x_1, \ldots, x_n$ (the real private chances for **B**!)

Prediction 3: **An emergent notion of objectivity**

Let's consider a colourful example:

A-world



Two probability distributions:

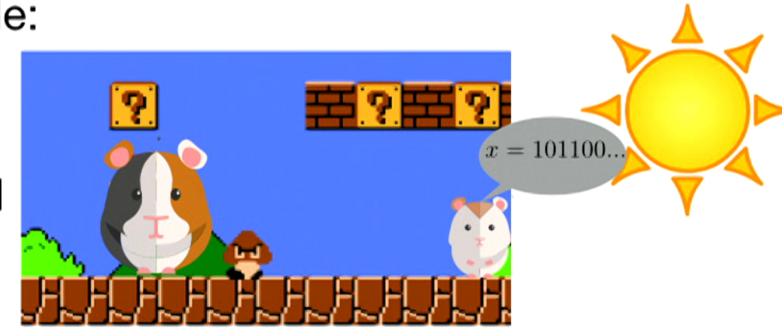
$\nu(x_1, x_2, \ldots, x_n) :=$ prob. that **B** is in states $x_1, \ldots, x_n$ acc. to **A**-world

$\mathbf{P}(x_1, \ldots, x_n) =$ algorithmic probability that **B** is in states $x_1, \ldots, x_n$
(the real private chances for **B**!)

Prediction 3: **An emergent notion of objectivity**

Let's consider a colourful example:

A-world



If Abby has a chance of about 100% of *seeing Bambi see the sun rise tomorrow*, then will Bambi have a chance of about 100% of seeing the sun rise tomorrow? ν $\mathbf{P}$

$\nu(x_1, x_2, \ldots, x_n) :=$ prob. that **B** is in states $x_1, \ldots, x_n$ acc. to **A**-world

$\mathbf{P}(x_1, \ldots, x_n) =$ algorithmic probability that **B** is in states $x_1, \ldots, x_n$ (the real private chances for **B**!)

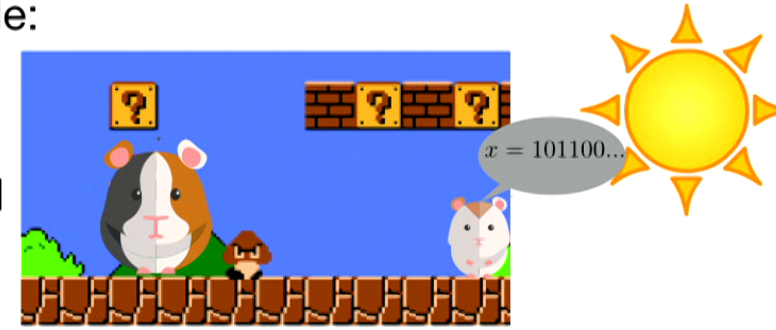
Surprise 1: Probabilistic zombies

- "Objective reality" is a theorem, not an assumption:
  $$\mathbf{P}(y|x_1,\ldots,x_k) \stackrel{k\to\infty}{\longrightarrow} \nu(y|x_1,\ldots,x_k).$$
  Sometimes premises of theorem not satisfied ⟶ **"zombies"!**

Prediction 3: **An emergent notion of objectivity**

Let's consider a colourful example:

A-world



If Abby has a chance of about 100% of *seeing Bambi see the sun rise tomorrow*, then will Bambi have a chance of about 100% of seeing the sun rise tomorrow? ν $\mathbf{P}$

Theorem: With ν-probability one,

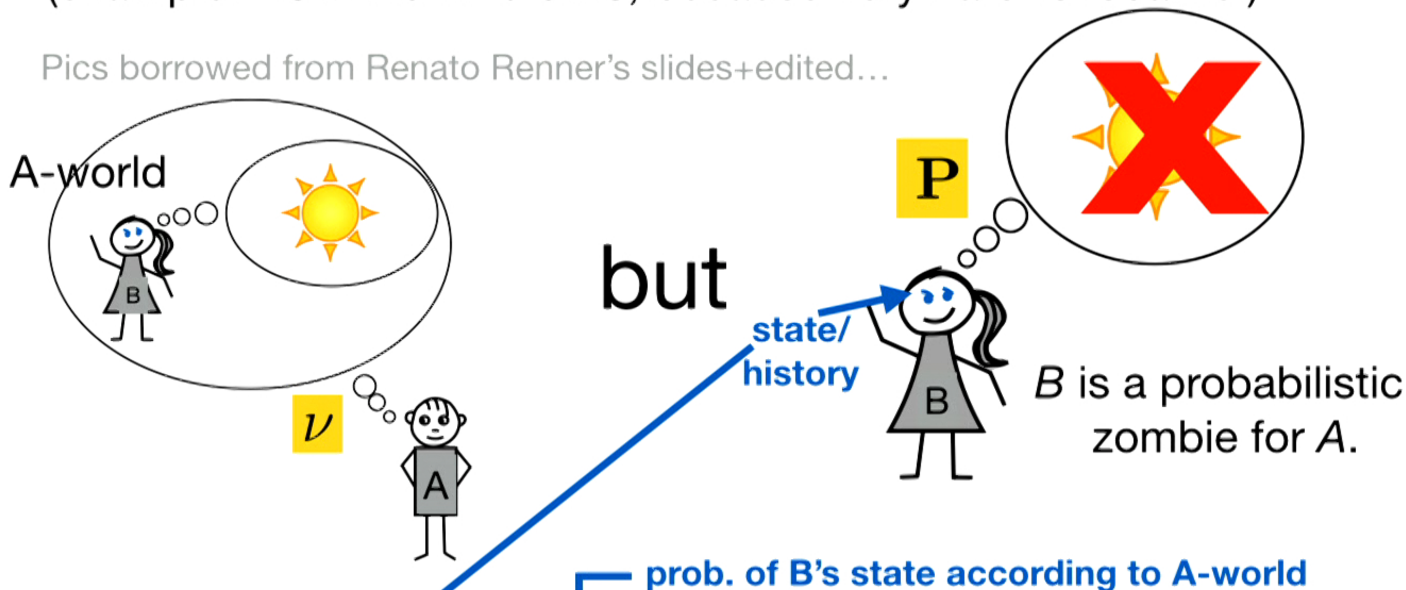
$$\mathbf{P}(y|x_1,\ldots,x_k) \stackrel{k\to\infty}{\longrightarrow} \nu(y|x_1,\ldots,x_k).$$

So the answer is "**yes**", asymptotically.

(In other words: $\mathbf{P} \approx \nu$ if **B** is "old enough" in **A**-world.)

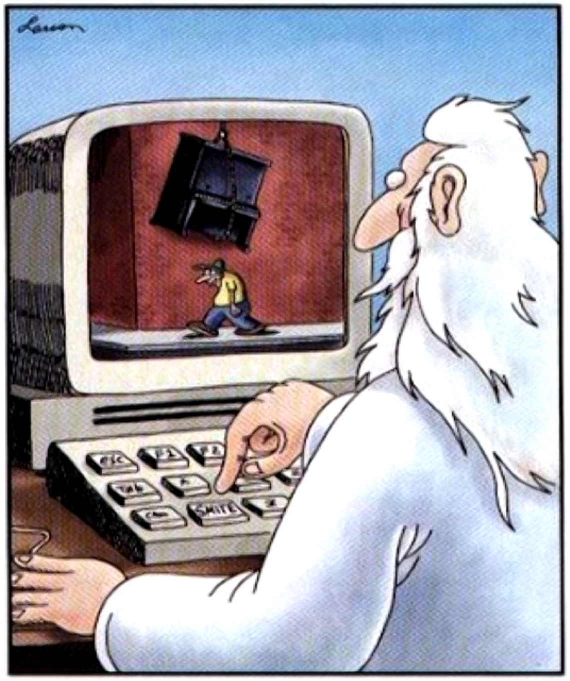
Surprise 1: Probabilistic zombies

- $K(\mathbf{x})$ too small: **A** “points to” something in his world that is too simple (e.g. a single bit, written on a blackboard)
- $K(\nu)$ too large: **A** “points to” something in a too complicated way (example: **Boltzmann brains**, because very hard to localize.)



Theorem: if $K(\mathbf{x}) \ll K(\nu)$ then zombie, i.e. $\mathbf{P} \neq \nu$.

Surprise 2: Brain emulation



God at His computer

Get also concrete criteria for when **simulation** of an agent corresponds to an “actual first-person perspective” (similarly as in the zombie case).

Turns out: makes big difference if simulation is **“open” or “closed”** (feed in outside data or not). More details in paper.

Conclusions

− Useless for most questions physicists care about.

O Proof of principle / **blueprint** of an "idealistic" predictive theory.

+ Many predictions / consequences from very simple assumptions.

- Existence of a simple computational probabilistic external world
- Emergence of objectivity (typically)
- Probabilistic zombies (in some cases)
- Resolves (versions of) the Boltzmann brain problem++
- No-signalling and Bell violation (**modulo an open problem**)
- Predictions for computer emulation of agents
- (Some sort of) subjective immortality, *but no possibility to use this for solving NP-complete problems in poly time*. (**But depends very much on details of the formulation.**)

Full version: **arXiv:1712.01826**
Short version (v2 soon): **arXiv:1712.01816**

4. Novel predictions

An open problem

An observer's **state** can be represented by a binary string (like $x_1 = 011010$). One (subjective) moment after the other, this yields a sequence $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, and the probability of the next state *y* is

$$\mathbf{P}(y|x_1, \ldots, x_n) := \frac{\mathbf{P}(x_1, \ldots, x_n, y)}{\mathbf{P}(x_1, \ldots, x_n)},$$

where **P** is conditional **algorithmic probability**.

Conceptually, it would be more consequential to define **P** only to depend on the present, not the past. In some sense, the "past" is only what an observer presently remembers…

$$\mathbf{P}(y|x_n).$$

Conceptually (much) clearer, but **consequences much harder to work out**. Don't know how to do it (yet).