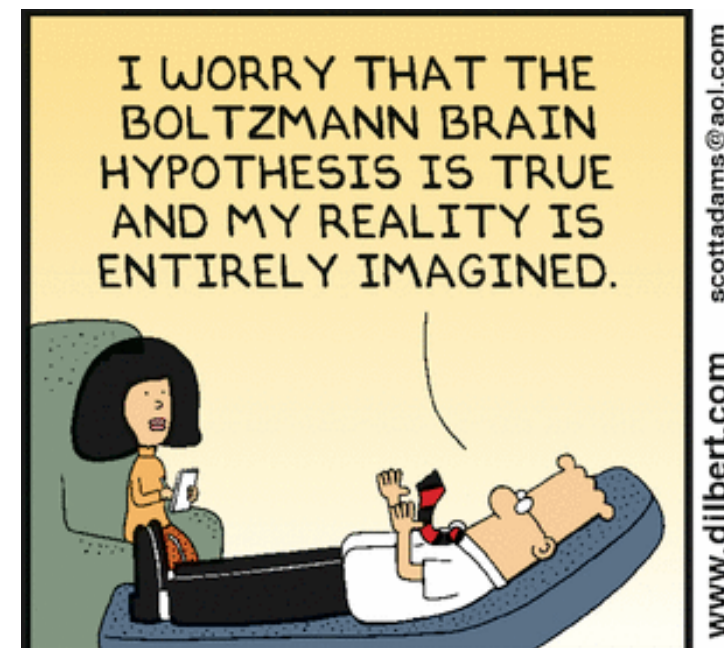
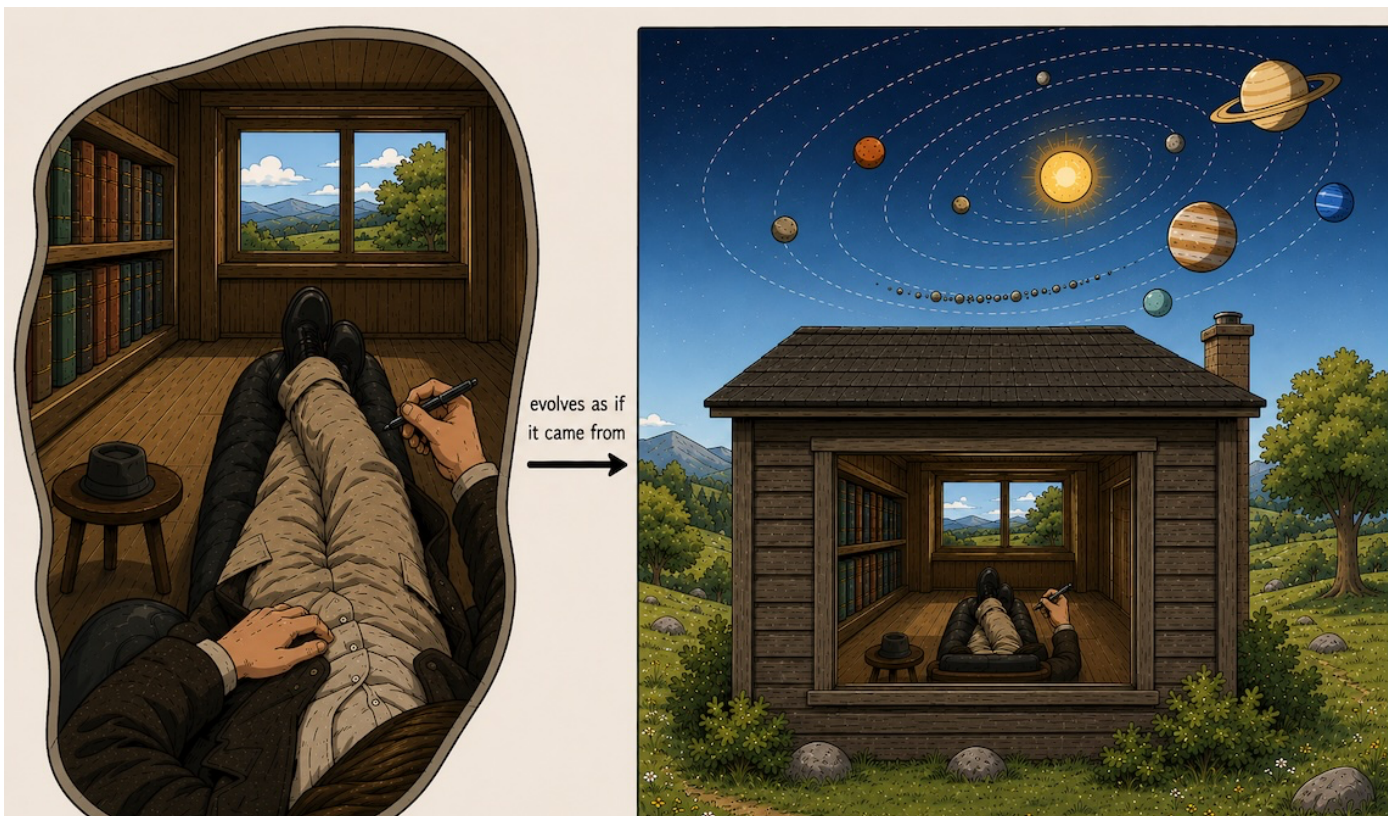


Algorithmic idealism: a first-person-first approach to physical reality

Markus P. Müller

Institute for Quantum Optics and Quantum Information (IQOQI), Vienna
Perimeter Institute for Theoretical Physics (PI), Waterloo, Canada



Müller Group @ IQOQI Vienna



Manuel Mekonnen (PhD student), **Markus Müller** (PI), **Caroline Jones** (PhD student), **Stefan Ludescher** (PhD student), **Thomas Galley** (postdoc), **Albert Aloy** (postdoc), **Jan Głowacki** (postdoc).

Quantum information and foundations of physics

Overview

1. Preparation: on **private experiments**
2. Limitations of the third-person view: **Restriction A**
3. A physicalist physician: why Restriction A for $n=1$ is bad
4. An overview on **Algorithmic Idealism**

Overview

1. Preparation: on **private experiments**
2. Limitations of the third-person view: **Restriction A**
3. A physicalist physician: why Restriction A for $n=1$ is bad
4. An overview on **Algorithmic Idealism**

Some regimes of scientific inquiry

Public experiments:

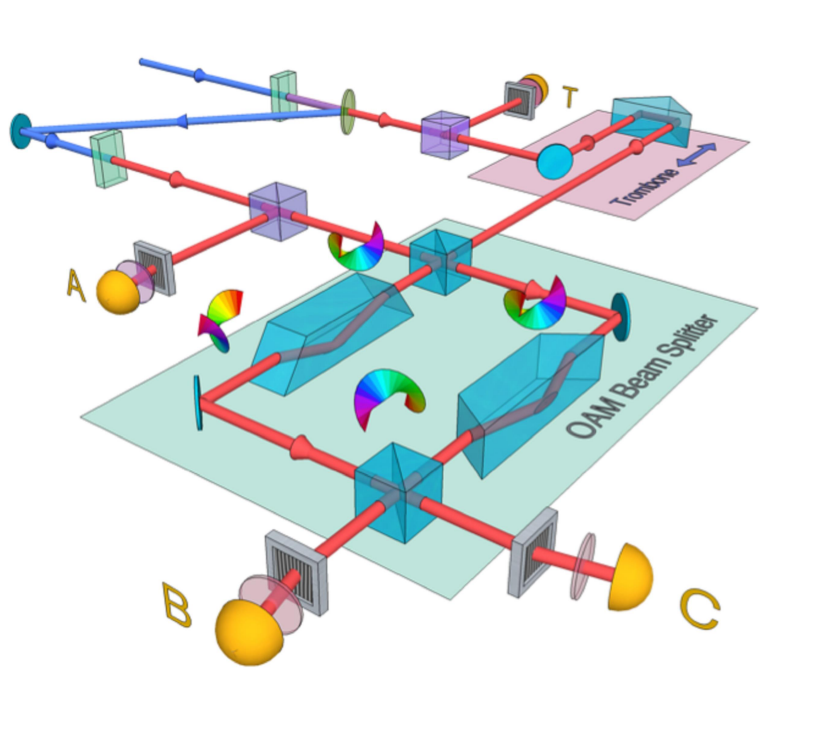
(Perhaps probabilistic) predictions for the outcomes of experiments that can be intersubjectively observed.

Some regimes of scientific inquiry

Public experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that can be intersubjectively observed.

“Gold standard” of scientific inquiry.



Credit: University of Vienna.



Credit: ESA.



Credit: NASA.

Some regimes of scientific inquiry

Ontological claims:

Answers to questions such as: What is really going on in the world? What is the world really made of?

Some regimes of scientific inquiry

Which interpretation of quantum theory is true?

Is there only a single world or a “multiverse”?

Did something happen before the Big Bang?

...

Ontological claims:

Answers to questions such as: What is really going on in the world? What is the world really made of?

Some regimes of scientific inquiry

Which interpretation of quantum theory is true?

Is there only a single world or a “multiverse”?

Did something happen before the Big Bang?

...

In this talk, this category of questions does not play any role.

(I think our intuition has evolved to help us survive and reproduce, but not to understand reality beyond its correlation with these goals.)

Ontological claims:

Answers to questions such as: What is really going on in the world? What is the world really made of?

Some regimes of scientific inquiry

Public experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that can be intersubjectively observed.

Ontological claims:

Answers to questions such as: What is really going on in the world? What is the world really made of?

Some regimes of scientific inquiry

Public experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that can be intersubjectively observed.

Ontological claims:

Answers to questions such as: What is really going on in the world? What is the world really made of?

Some regimes of scientific inquiry

Public experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that can be intersubjectively observed.

Private experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that can (only?) be privately observed.

Ontological claims:

Answers to questions such as: What is really going on in the world? What is the world really made of?

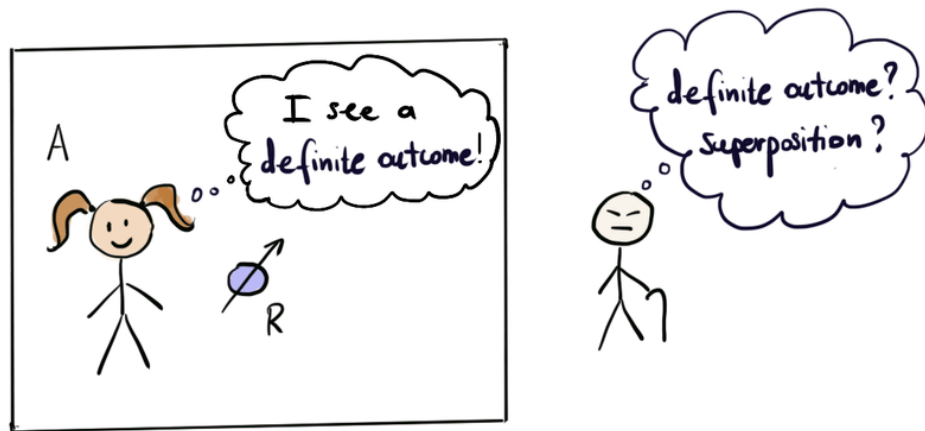
Some regimes of scientific inquiry

Private experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that can (only?) be privately observed.

Some regimes of scientific inquiry

Wigner's Friend



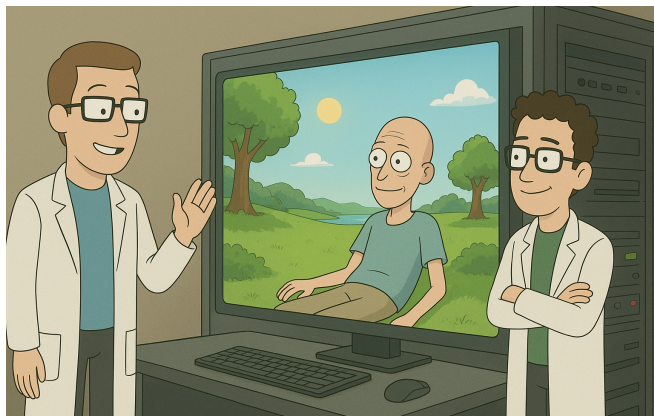
Boltzmann brain problem



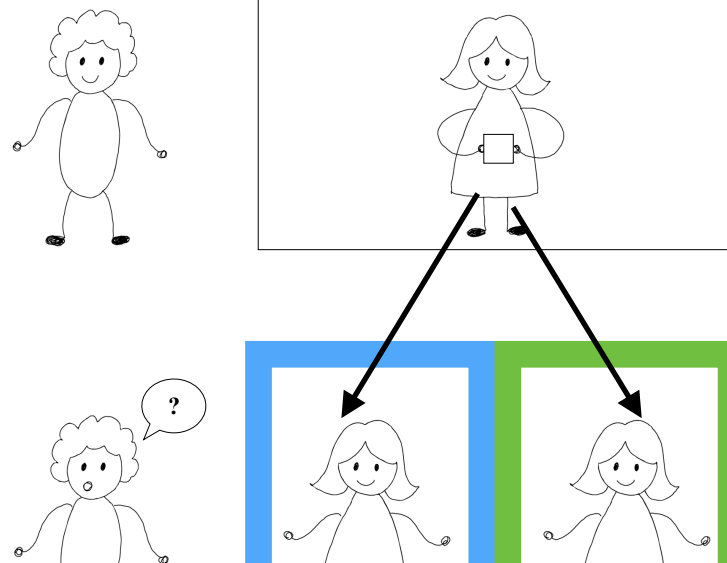
Private experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that can (only?) be privately observed.

Computer simulation of agents



Duplication



Death



Some regimes of scientific inquiry

Non-example 1: Qualia.

“Does *my* red feel like *your* red?”

My considerations do not apply to qualia; only “speakable” observations.

Private experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that can (only?) be privately observed.

Some regimes of scientific inquiry

Non-example 1: Qualia.

“Does *my* red feel like *your* red?”

My considerations do not apply to qualia; only “speakable” observations.

Private experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that can (only?) be privately observed.

Non-example 2: (The hard problem of) consciousness.

While this points at a conceptually similar difficulty (private experience that cannot directly be assessed externally), it is more of the “qualia kind”, hence not relevant here.

Some regimes of scientific inquiry

Public experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that can be intersubjectively observed.

Private experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that can (only?) be privately observed.

Some regimes of scientific inquiry

Public experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that can be intersubjectively observed.

Private experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that can ~~(only?)~~ be privately observed.

Some regimes of scientific inquiry

Public experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that can be intersubjectively observed.

Private experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that can ~~(only?)~~ be privately observed.

If outcomes can be intersubjectively observed, then they can be privately observed by *somebody*.

Hence, considering “private experiments” can be regarded as a natural **extension** of the regime of interest.

Some regimes of scientific inquiry

(Not necessarily public) experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that *somebody* can observe.

“What should I expect to observe / experience?”

Some regimes of scientific inquiry

(Not necessarily public) experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that *somebody* can observe.

“What should I expect to observe / experience?”

Quantum theory is arguably formulated (and certainly *used*) in this way.

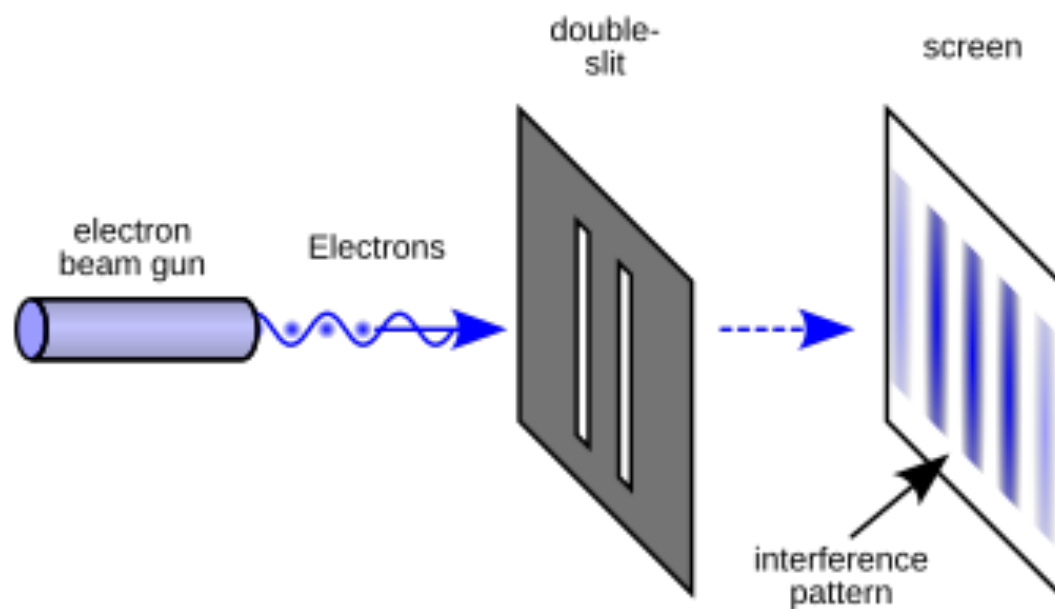
Some regimes of scientific inquiry

(Not necessarily public) experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that *somebody* can observe.

“What should I expect to observe / experience?”

Quantum theory is arguably formulated (and certainly *used*) in this way.



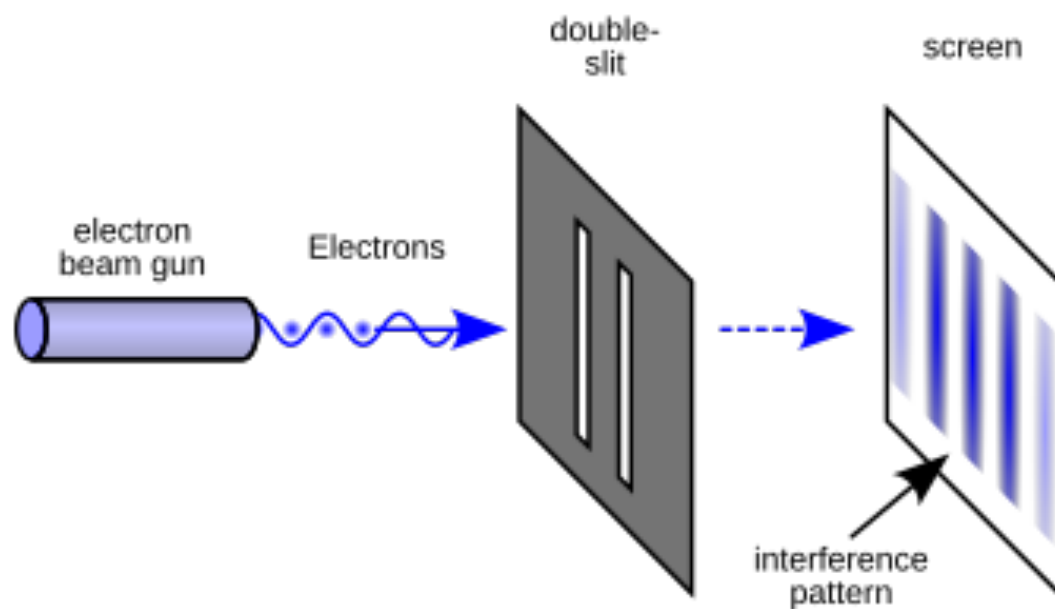
Some regimes of scientific inquiry

(Not necessarily public) experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that *somebody* can observe.

“What should I expect to observe / experience?”

Quantum theory is arguably formulated (and certainly *used*) in this way.



“Will the particle pass through the left or the right slit?”

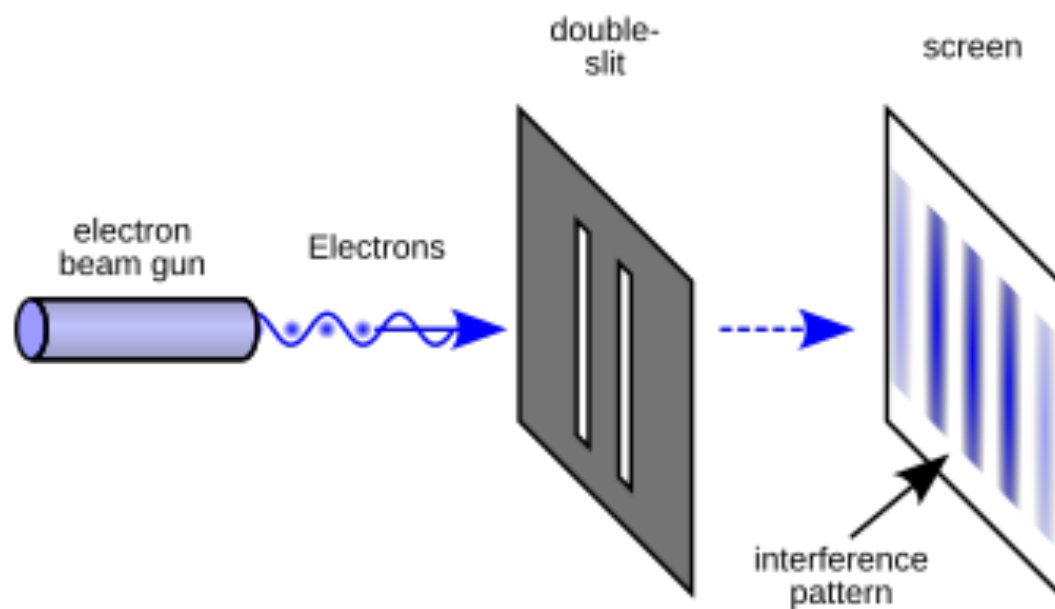
Some regimes of scientific inquiry

(Not necessarily public) experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that *somebody* can observe.

“What should I expect to observe / experience?”

Quantum theory is arguably formulated (and certainly *used*) in this way.



“Will the particle pass through the left or the right slit?” ✗

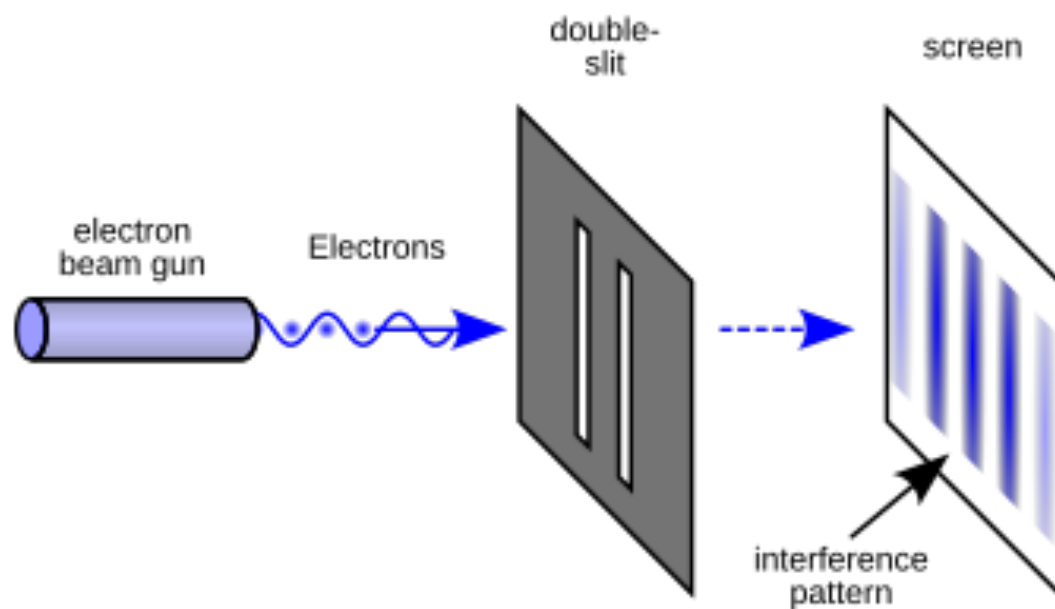
Some regimes of scientific inquiry

(Not necessarily public) experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that *somebody* can observe.

“What should I expect to observe / experience?”

Quantum theory is arguably formulated (and certainly *used*) in this way.



“Will the particle pass through the left or the right slit?” ✗

“Will I see a spot on the final screen in region R or not?”

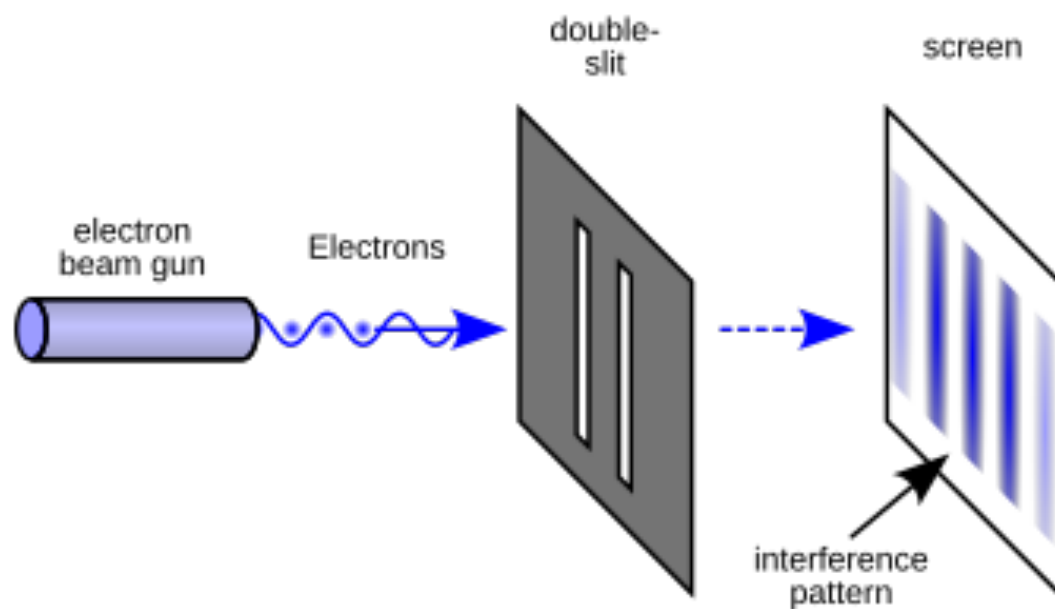
Some regimes of scientific inquiry

(Not necessarily public) experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that *somebody* can observe.

“What should I expect to observe / experience?”

Quantum theory is arguably formulated (and certainly *used*) in this way.



“Will the particle pass through the left or the right slit?” ✗

“Will I see a spot on the final screen in region R or not?” ✓

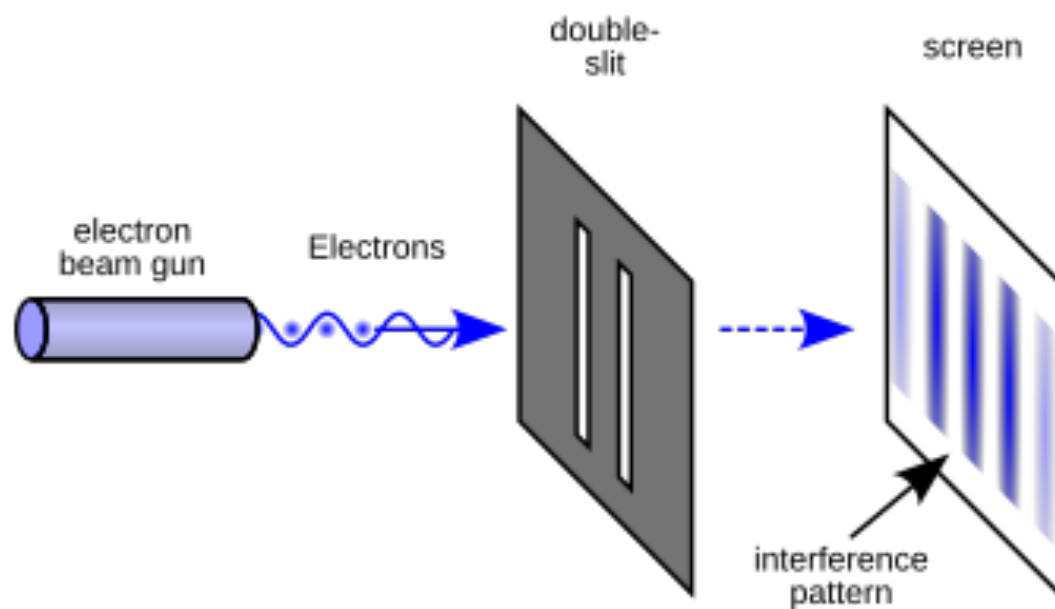
Some regimes of scientific inquiry

(Not necessarily public) experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that *somebody* can observe.

“What should I expect to observe / experience?”

Quantum theory is arguably formulated (and certainly *used*) in this way.



“Will the particle pass through the left or the right slit?” ✗

“Will I see a spot on the final screen in region R or not?” ✓

(Probabilistic) predictions only for **the latter**, not for **the former**.

Overview

1. Preparation: on **private experiments**
2. Limitations of the third-person view: **Restriction A**
3. A physicalist physician: why Restriction A for $n=1$ is bad
4. An overview on **Algorithmic Idealism**

Overview

1. Preparation: on **private experiments**

2. Limitations of the third-person view: **Restriction A**

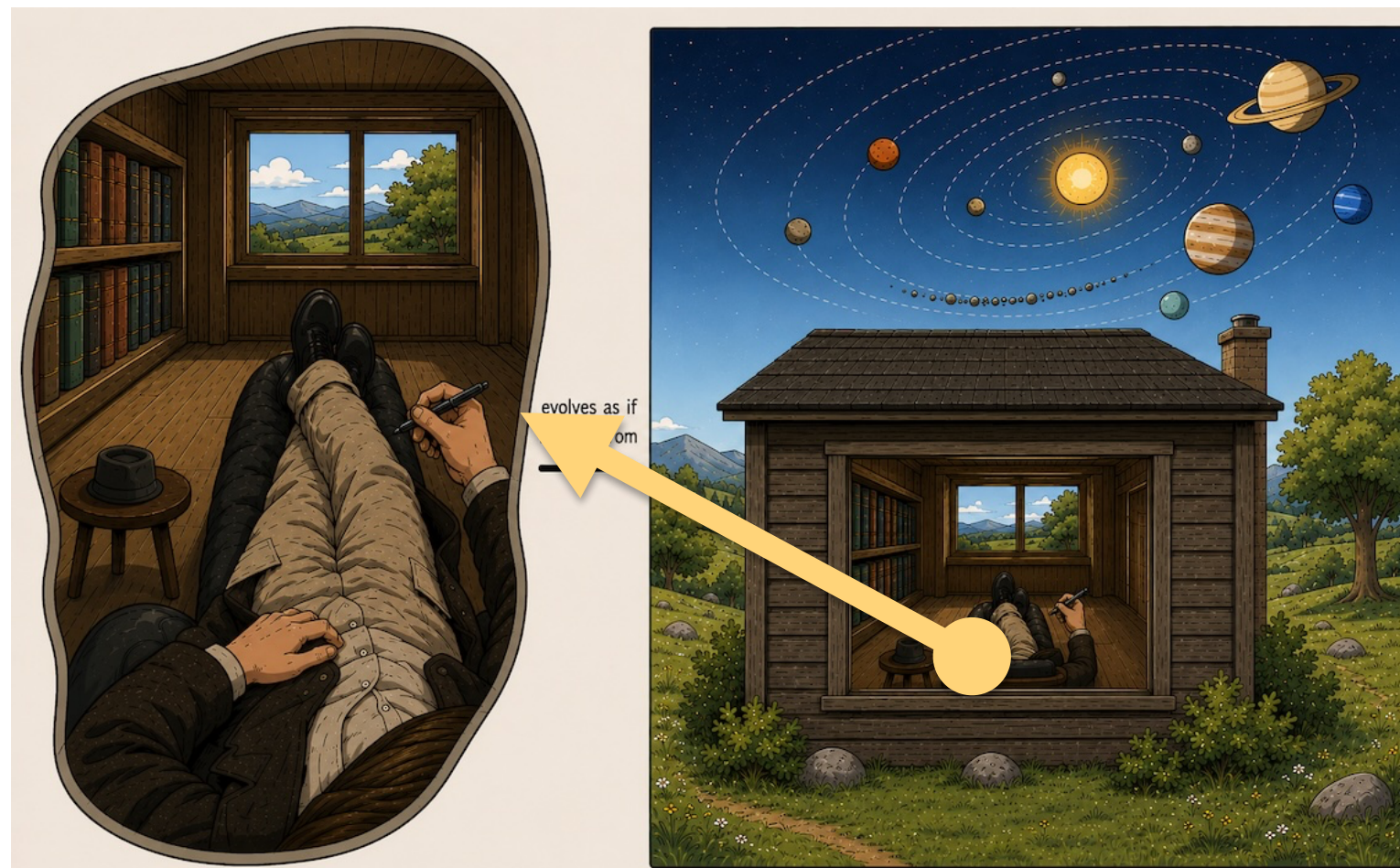
3. A physicalist physician: why Restriction A for $n=1$ is bad

4. An overview on **Algorithmic Idealism**

Limitations of the third-person view: **Restriction A**

To "what will I observe?", we typically use a **standard method**:

Determine your location in the world; model the evolution of the universe; from the (probabilities of the) facts of the world, take your local marginal.

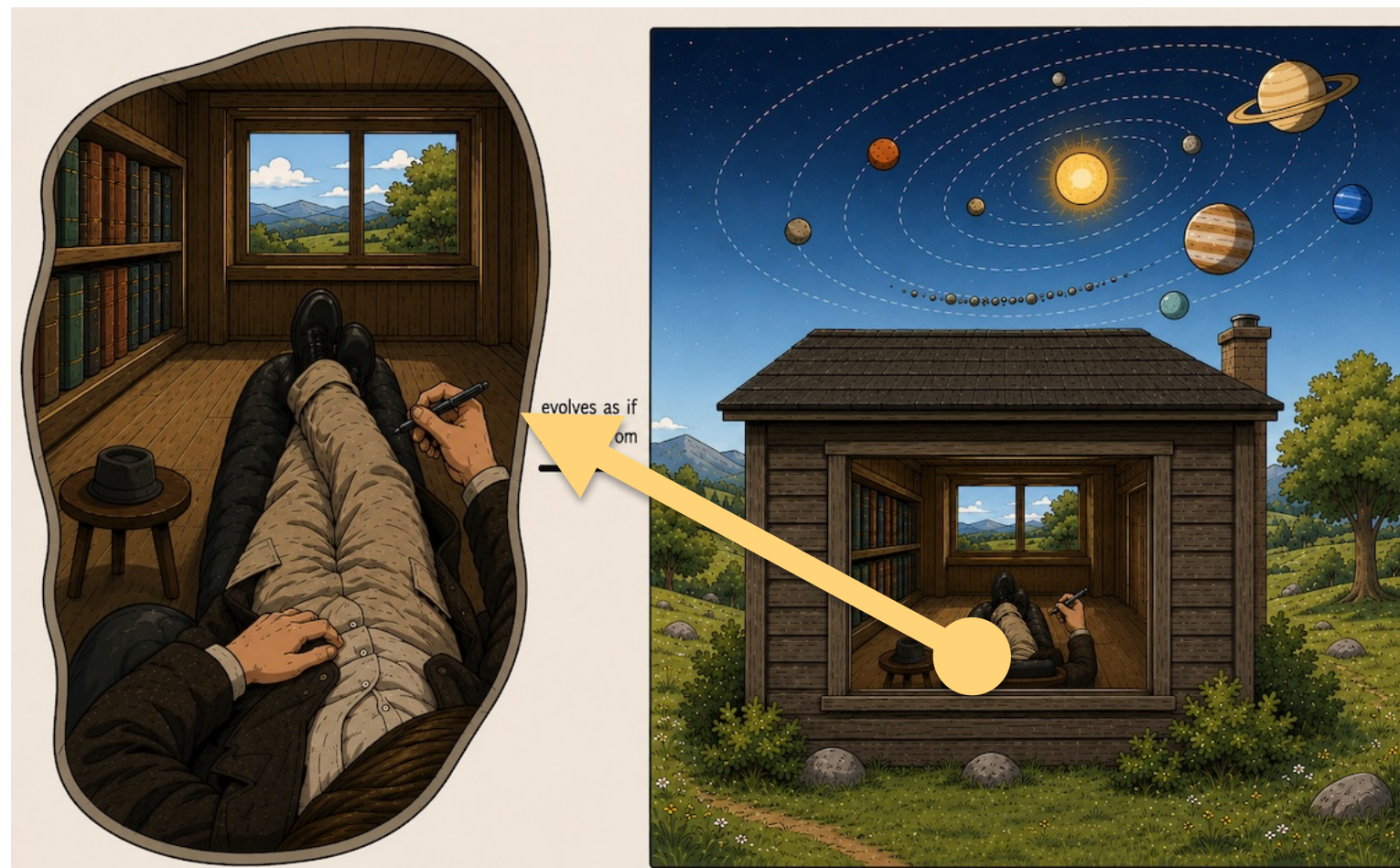


Limitations of the third-person view: **Restriction A**

To "what will I observe?", we typically use a **standard method**:

Determine your location in the world; model the evolution of the universe; from the (probabilities of the) facts of the world, take your local marginal.

However, this method is not always available:



Limitations of the third-person view: **Restriction A**

To "what will I observe?", we typically use a **standard method**:

Determine your location in the world; model the evolution of the universe; from the (probabilities of the) facts of the world, take your local marginal.

However, this method is not always available:

Restriction A: For some experiments, our physical theories do not / cannot give us joint distributions for the future observations of all (n) agents.

Limitations of the third-person view: **Restriction A**

To "what will I observe?", we typically use a **standard method**:

Determine your location in the world; model the evolution of the universe; from the (probabilities of the) facts of the world, take your local marginal.

However, this method is not always available:

Restriction A: For some experiments, our physical theories do not / cannot give us joint distributions for the future observations of all (n) agents.

This means different things for the following two cases:

Limitations of the third-person view: **Restriction A**

To "what will I observe?", we typically use a **standard method**:

Determine your location in the world; model the evolution of the universe; from the (probabilities of the) facts of the world, take your local marginal.

However, this method is not always available:

Restriction A: For some experiments, our physical theories do not / cannot give us joint distributions for the future observations of all (n) agents.

This means different things for the following two cases:

- **$n=1$:** The agent cannot use its physical theory to predict what it should (privately) expect to observe in the experiment.

Limitations of the third-person view: **Restriction A**

To "what will I observe?", we typically use a **standard method**:

Determine your location in the world; model the evolution of the universe; from the (probabilities of the) facts of the world, take your local marginal.

However, this method is not always available:

Restriction A: For some experiments, our physical theories do not / cannot give us joint distributions for the future observations of all (n) agents.

This means different things for the following two cases:

- **$n=1$:** The agent cannot use its physical theory to predict what it should (privately) expect to observe in the experiment.
- **$n \geq 2$:** Perhaps every single agent can obtain predictions for what it will observe, but these individual probabilistic predictions **cannot** be deduced (say, as marginals) from a **joint** distribution of "facts of the world".

Example: extended **Wigner's friend** scenarios

nature
physics

ARTICLES

<https://doi.org/10.1038/s41567-020-0990-x>



A strong no-go theorem on the Wigner's friend paradox

Kok-Wei Bong^{1,4}, Aníbal Utreras-Alarcón^{1,4}, Farzad Ghafari ¹, Yeong-Cherng Liang²,
Nora Tischler ¹✉, Eric G. Cavalcanti ³✉, Geoff J. Pryde ¹ and Howard M. Wiseman ¹

Example: extended Wigner's friend scenarios

nature
physics

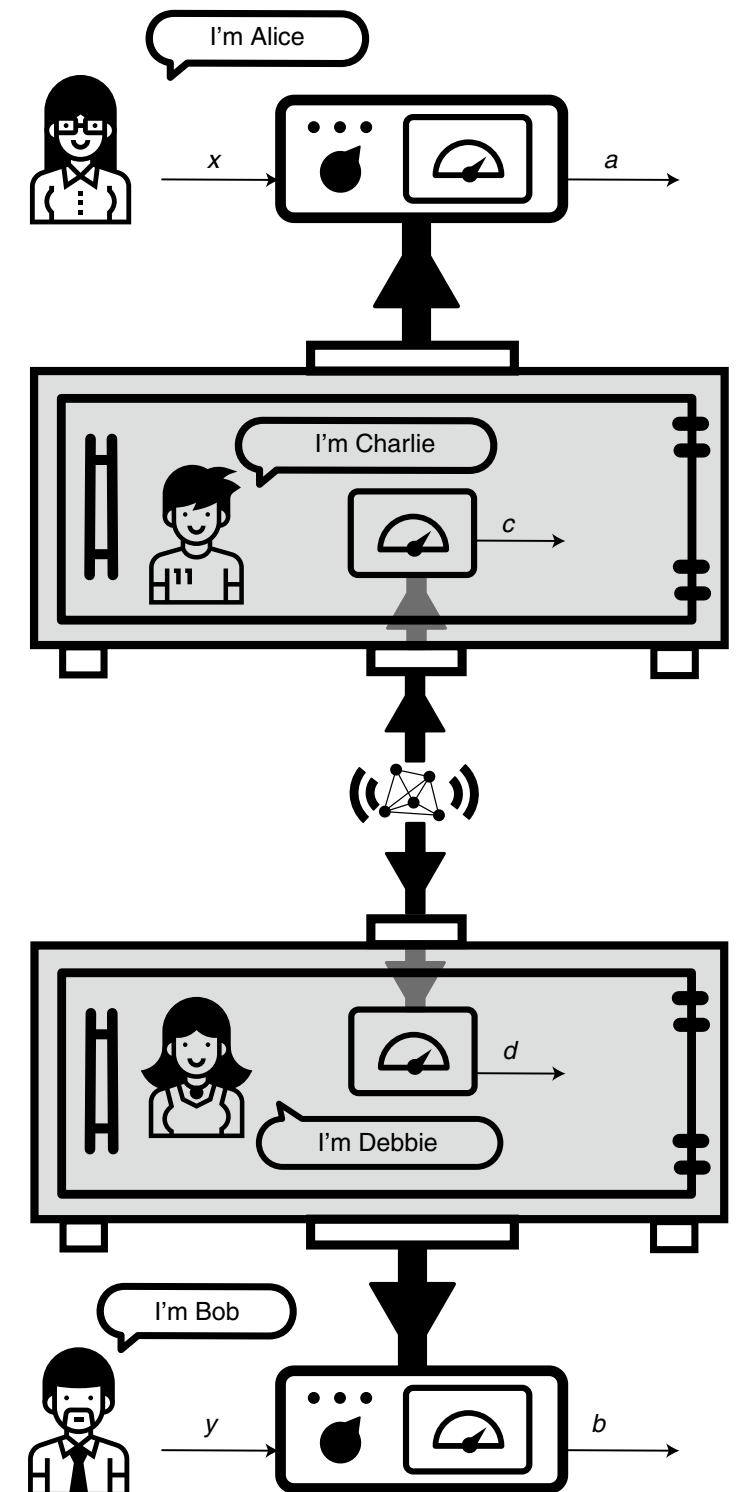
ARTICLES

<https://doi.org/10.1038/s41567-020-0990-x>



A strong no-go theorem on the Wigner's friend paradox

Kok-Wei Bong^{1,4}, Aníbal Utreras-Alarcón^{1,4}, Farzad Ghafari¹, Yeong-Cherng Liang²,
Nora Tischler¹, Eric G. Cavalcanti³, Geoff J. Pryde¹ and Howard M. Wiseman¹



If $x=1$, Alice asks Charlie for its outcome c and outputs $a=c$. (Debbie can be dropped.)

Example: extended **Wigner's friend** scenarios

nature
physics

ARTICLES

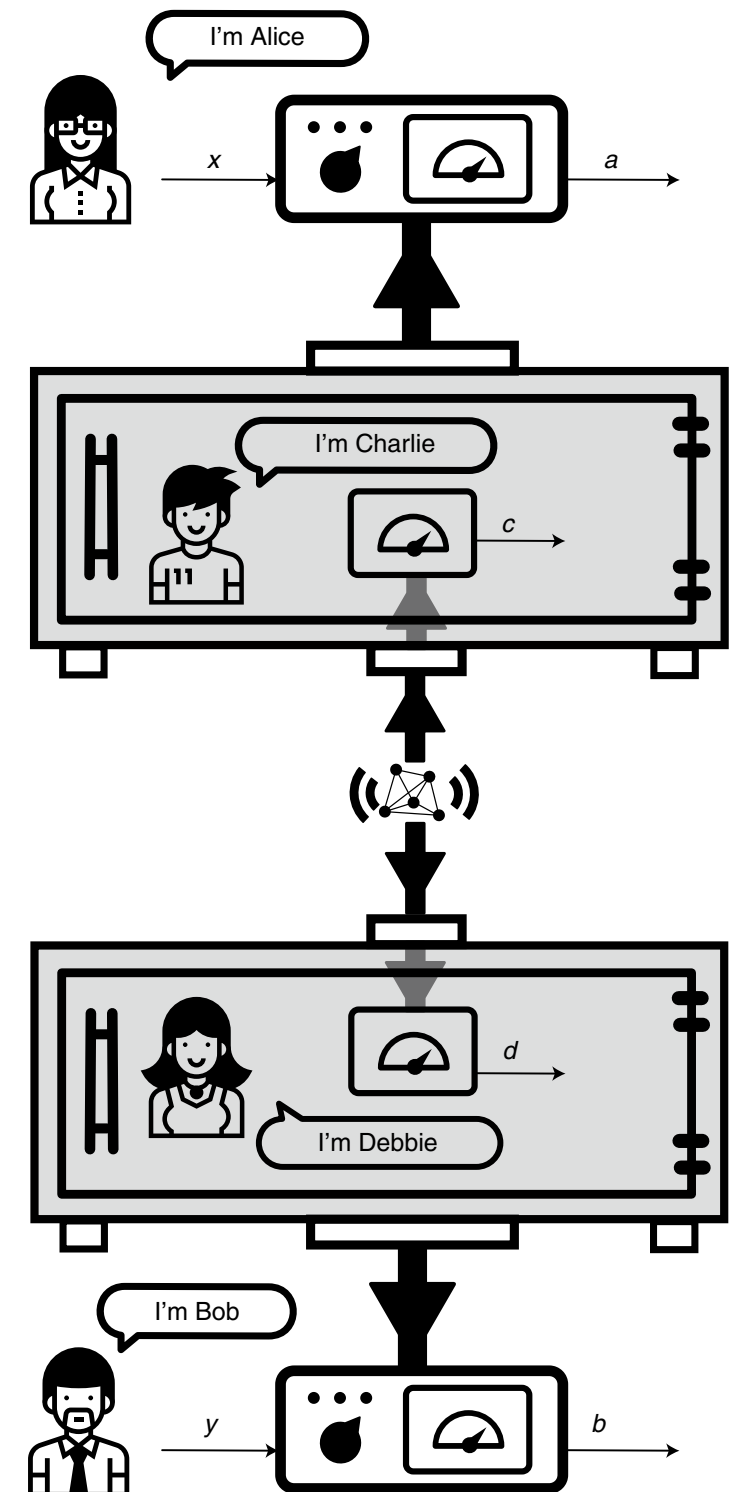
<https://doi.org/10.1038/s41567-020-0990-x>

Check for updates

A strong no-go theorem on the Wigner's friend paradox

Kok-Wei Bong^{1,4}, Aníbal Utreras-Alarcón^{1,4}, Farzad Ghafari¹, Yeong-Cherng Liang², Nora Tischler¹, Eric G. Cavalcanti³, Geoff J. Pryde¹ and Howard M. Wiseman¹

If the statistics violates a so-called “local friendliness inequality”, then the following three propositions cannot all be true:



If $x=1$, Alice asks Charlie for its outcome c and outputs $a=c$.

Example: extended Wigner's friend scenarios

nature
physics

ARTICLES

<https://doi.org/10.1038/s41567-020-0990-x>

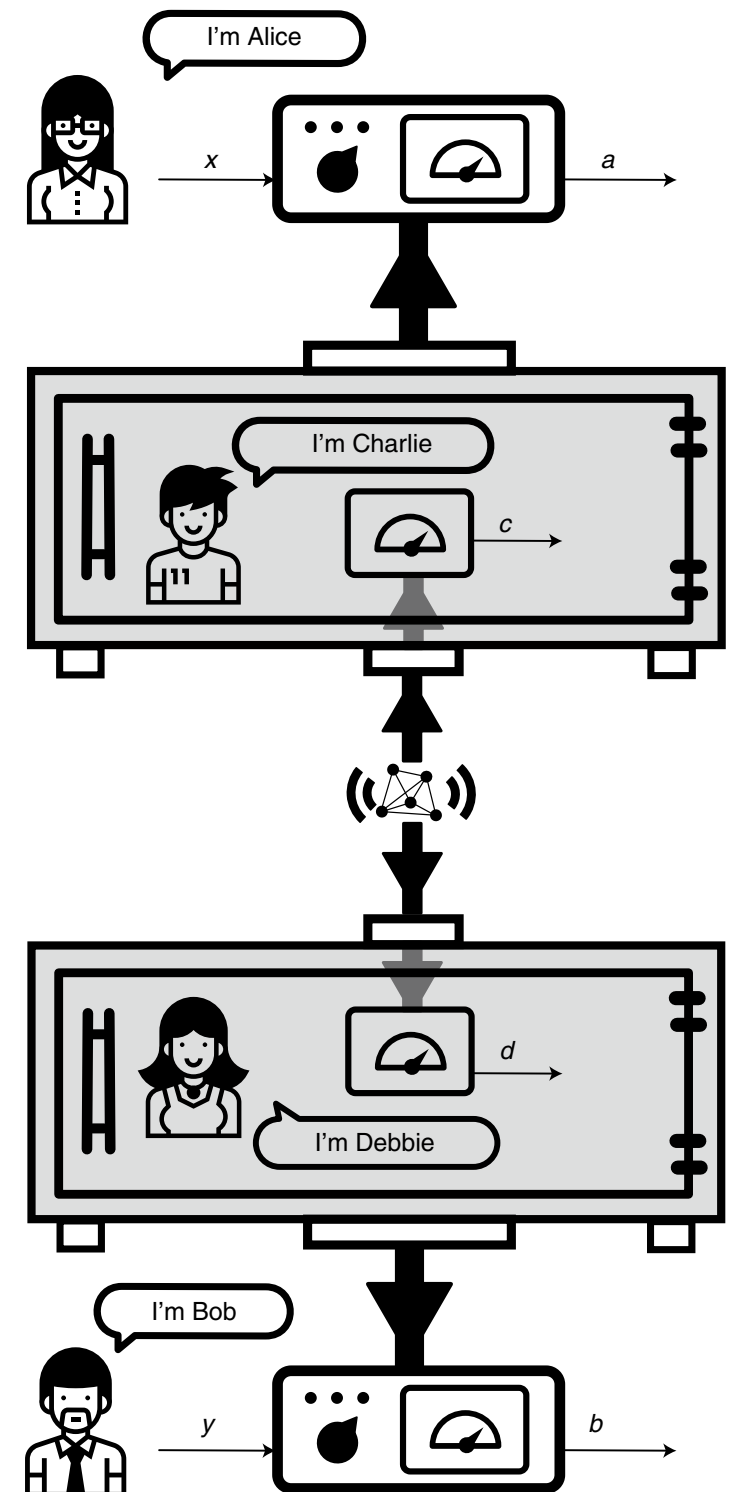
Check for updates

A strong no-go theorem on the Wigner's friend paradox

Kok-Wei Bong^{1,4}, Aníbal Utreras-Alarcón^{1,4}, Farzad Ghafari¹, Yeong-Cherng Liang², Nora Tischler¹, Eric G. Cavalcanti³, Geoff J. Pryde¹ and Howard M. Wiseman¹

If the statistics violates a so-called “local friendliness inequality”, then the following three propositions cannot all be true:

Locality, No Superdeterminism,
Absoluteness of Observed Events:



If $x=1$, Alice asks Charlie for its outcome c and outputs $a=c$.

Example: extended Wigner's friend scenarios

nature
physics

ARTICLES

<https://doi.org/10.1038/s41567-020-0990-x>



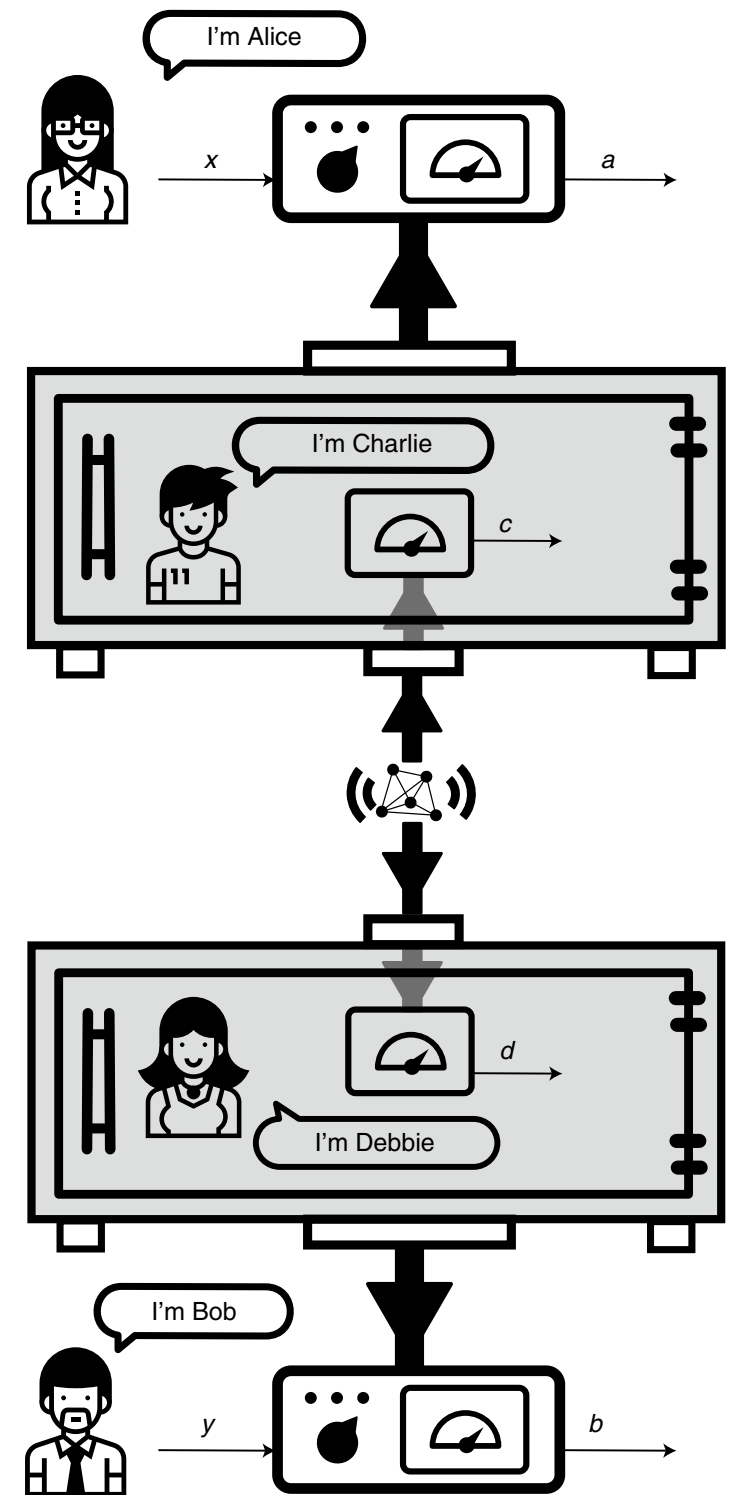
A strong no-go theorem on the Wigner's friend paradox

Kok-Wei Bong^{1,4}, Aníbal Utreras-Alarcón^{1,4}, Farzad Ghafari¹, Yeong-Cherng Liang², Nora Tischler¹, Eric G. Cavalcanti³, Geoff J. Pryde¹ and Howard M. Wiseman¹

If the statistics violates a so-called “local friendliness inequality”, then the following three propositions cannot all be true:

Locality, No Superdeterminism,
Absoluteness of Observed Events:

For every x, y , there is a joint prob. distr. $P(a, b, c, d | x, y)$ reproducing the observed distribution $p(a, b | x, y)$ as its marginal.



If $x=1$, Alice asks Charlie for its outcome c and outputs $a=c$.

Example: extended Wigner's friend scenarios

nature
physics

ARTICLES

<https://doi.org/10.1038/s41567-020-0990-x>

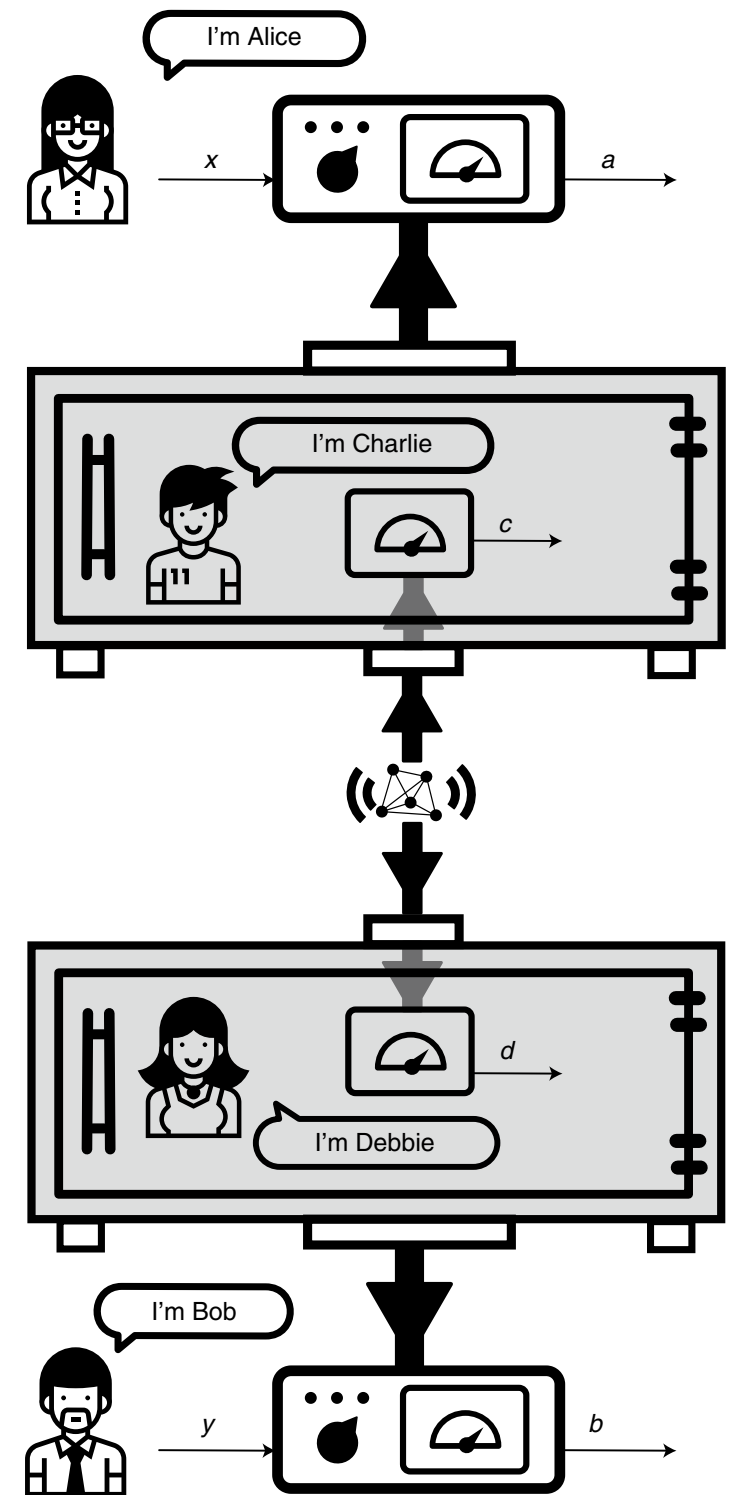
Check for updates

A strong no-go theorem on the Wigner's friend paradox

Kok-Wei Bong^{1,4}, Aníbal Utreras-Alarcón^{1,4}, Farzad Ghafari¹, Yeong-Cherng Liang², Nora Tischler¹, Eric G. Cavalcanti³, Geoff J. Pryde¹ and Howard M. Wiseman¹

If the statistics violates a so-called “local friendliness inequality”, then the following three propositions cannot all be true: [...]

Hence, **every** signal-local theory without superdeterminism, describing the experiment adequately as predicted by quantum theory, suffers from **Restriction A** for $n=4$ agents.



If $x=1$, Alice asks Charlie for its outcome c and outputs $a=c$. (Debbie can be dropped.)

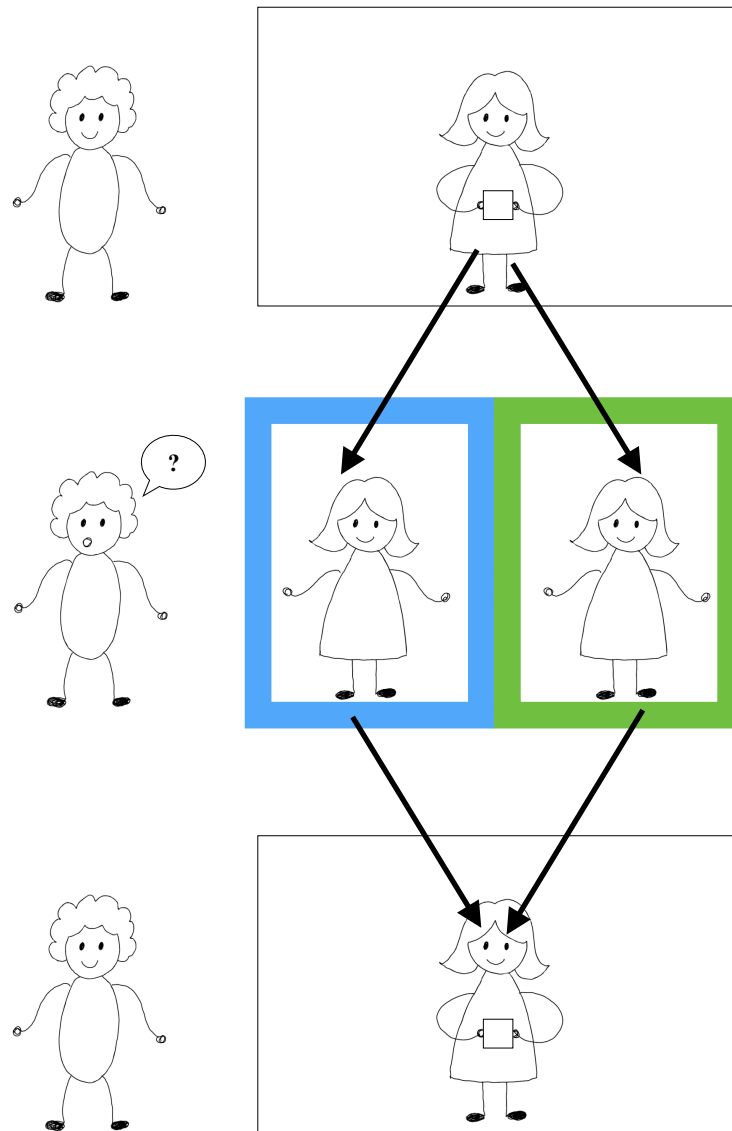
Duplication and (perhaps) fusion: **Restriction A** for $n=1$

Duplication and (perhaps) fusion: **Restriction A** for $n=1$

Suppose Freya is a human, and we have science fiction technology (or perhaps Freya is an AI, and we have *today's* technology).

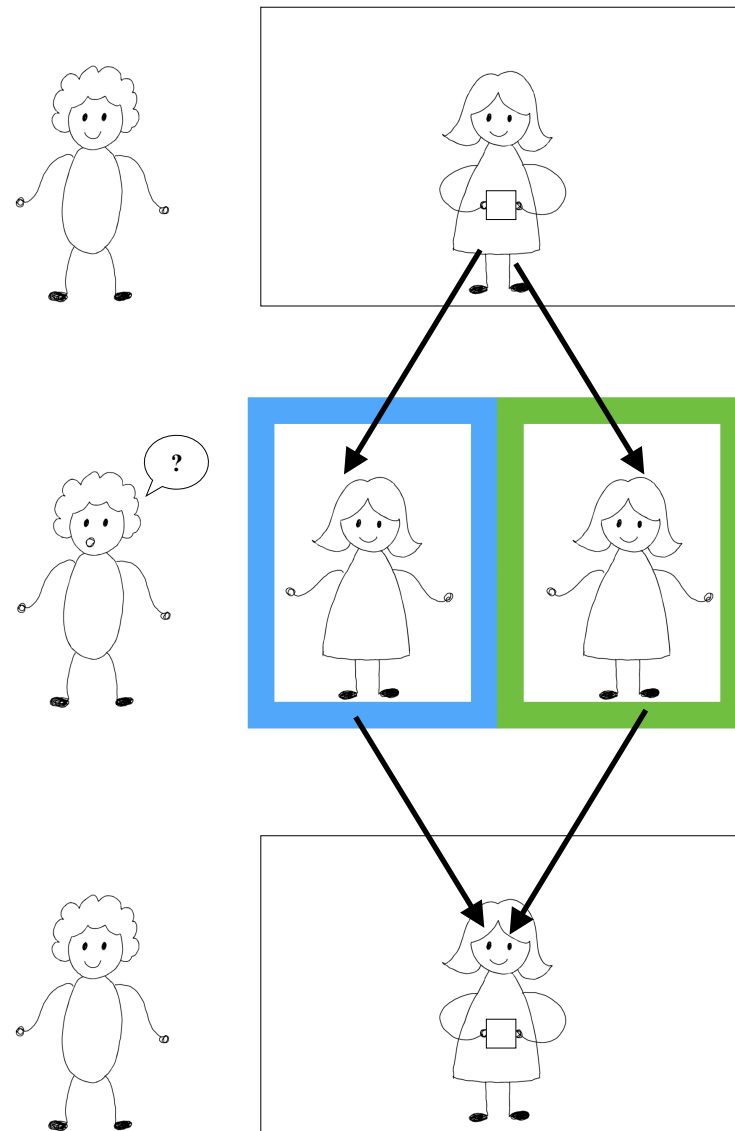
Duplication and (perhaps) fusion: **Restriction A** for $n=1$

Suppose Freya is a human, and we have science fiction technology (or perhaps Freya is an AI, and we have *today's* technology).



Duplication and (perhaps) fusion: **Restriction A** for $n=1$

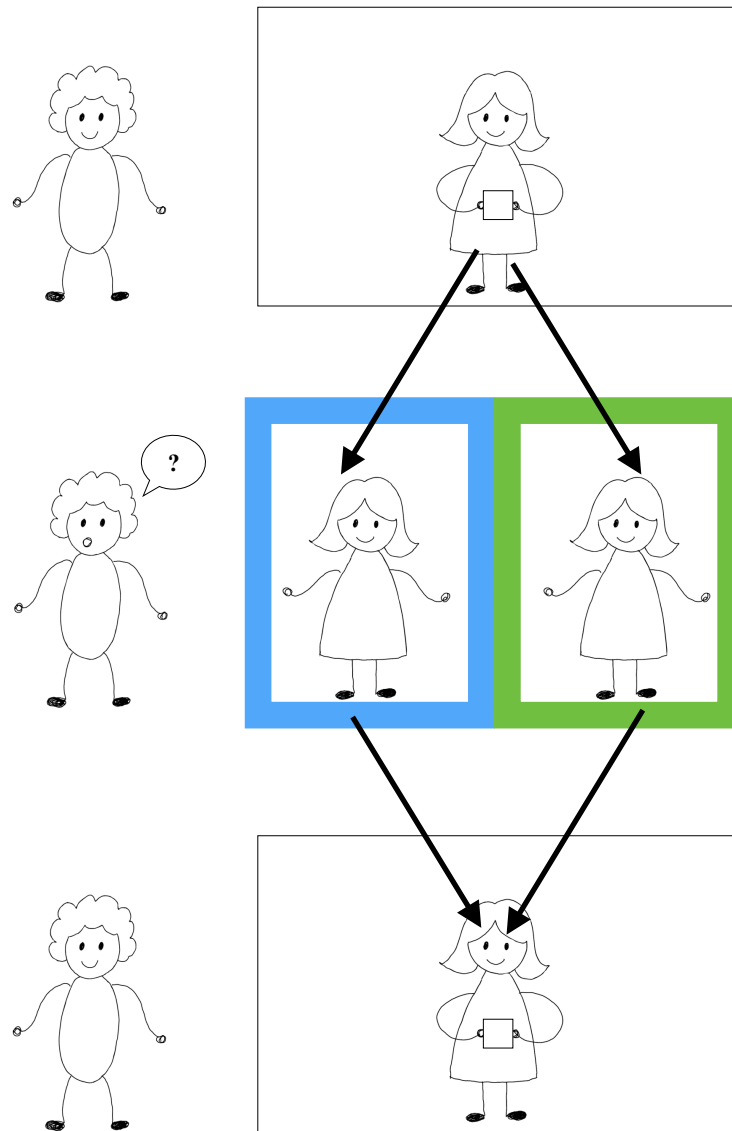
Suppose Freya is a human, and we have science fiction technology (or perhaps Freya is an AI, and we have *today's* technology).



Classical statistical physics has nothing at all to say about what Freya should believe about the color of room she will see.

Duplication and (perhaps) fusion: **Restriction A** for $n=1$

Suppose Freya is a human, and we have science fiction technology (or perhaps Freya is an AI, and we have *today's* technology).

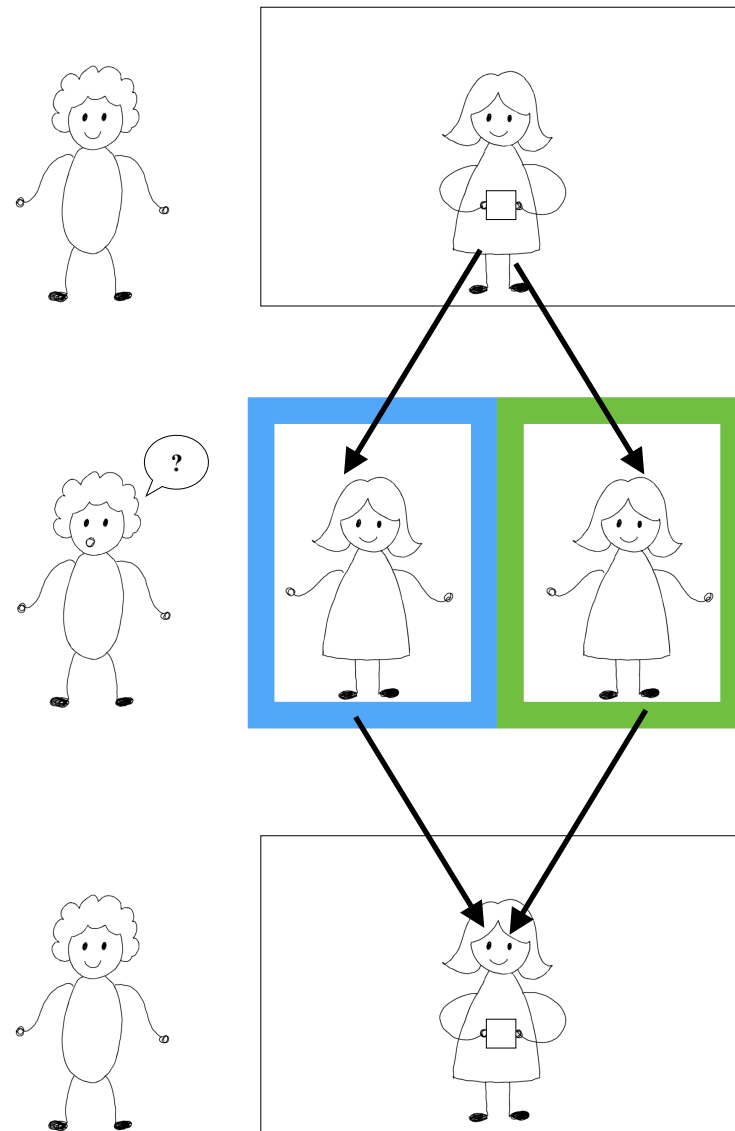


Classical statistical physics has nothing at all to say about what Freya should believe about the color of room she will see.

This is because all claims of classical physics are **intersubjectively testable** (including probabilistic claims), but not any claim of the form $P_F(\text{blue}) = p$: no external observer can test this claim statistically or otherwise.

Duplication and (perhaps) fusion: **Restriction A** for $n=1$

Suppose Freya is a human, and we have science fiction technology (or perhaps Freya is an AI, and we have *today's* technology).



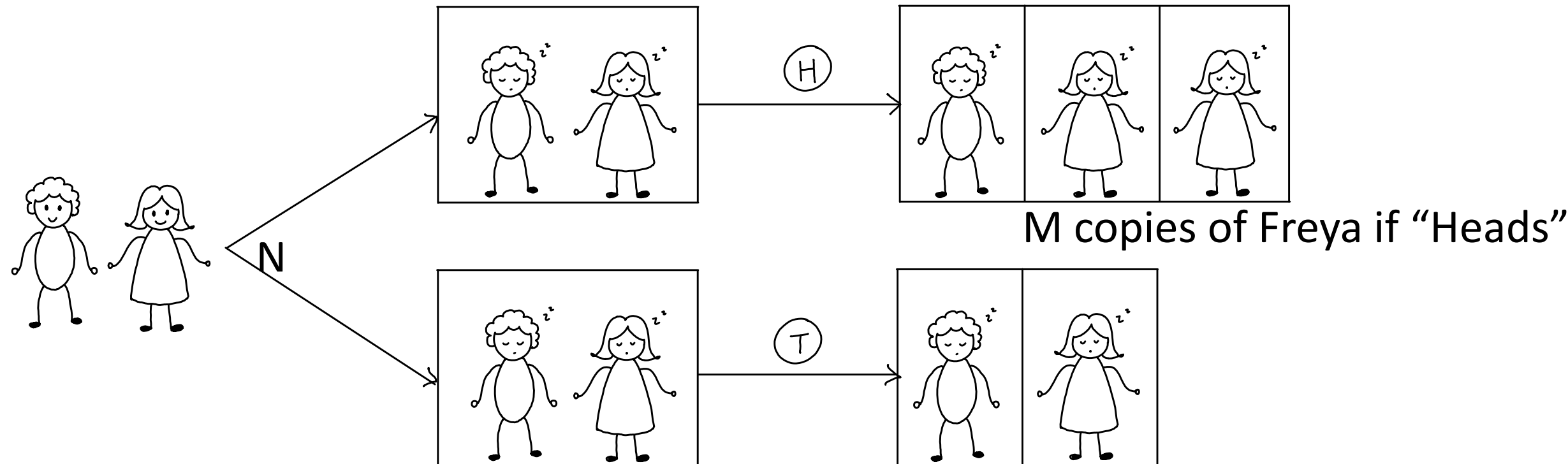
Classical statistical physics has nothing at all to say about what Freya should believe about the color of room she will see.

This is because all claims of classical physics are **intersubjectively testable** (including probabilistic claims), but not any claim of the form $P_F(\text{blue}) = p$: no external observer can test this claim statistically or otherwise.

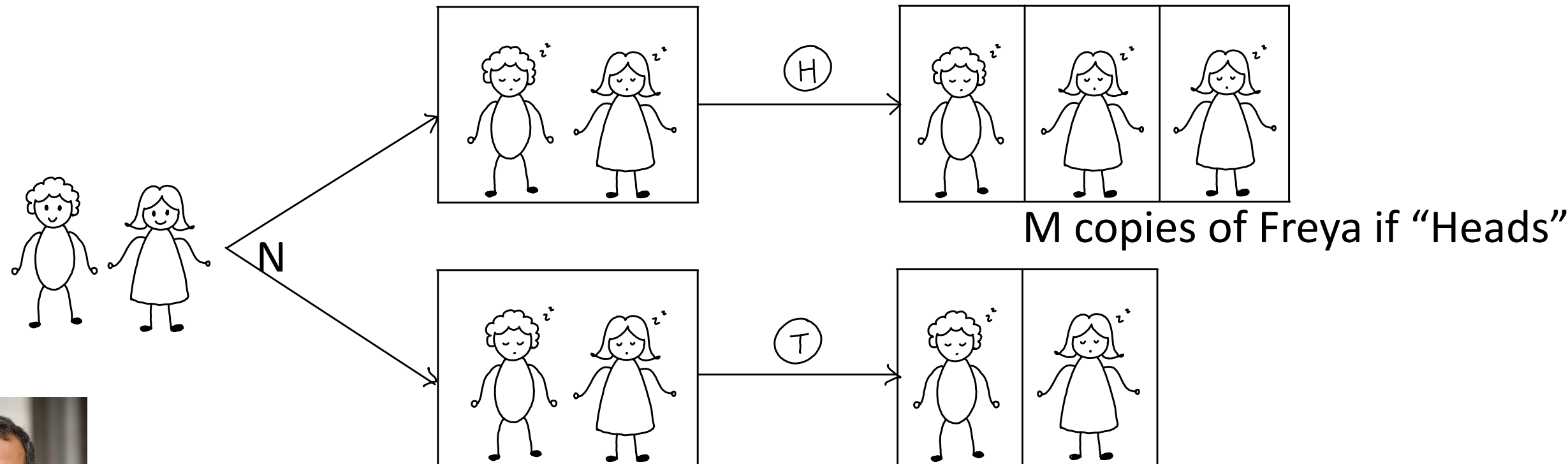
Hence, **Restriction A** applies to classical (statistical) physics for this scenario, for $n=1$ observer.

Quick advertisement: many more **interesting non-quantum examples**

Quick advertisement: many more **interesting non-quantum** examples



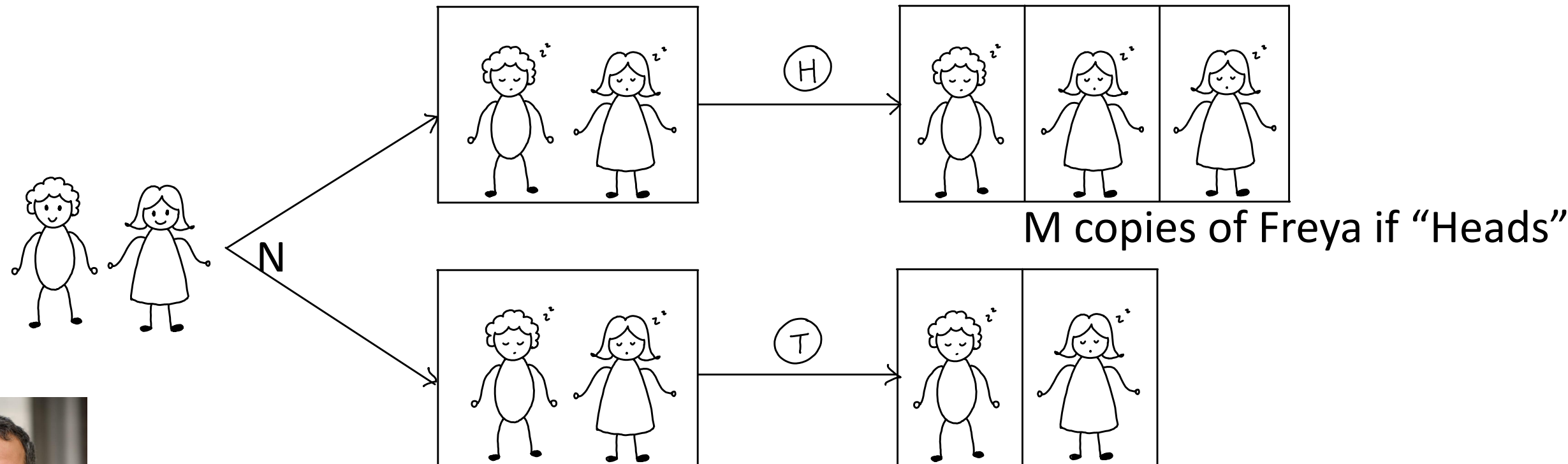
Quick advertisement: many more interesting non-quantum examples



Adam Elga (Princeton)

- **Elga's Principle of Indifference:**
Similar centered worlds deserve equal credence.
- **Principle CP (probabilistic consistency)**
- **NUI (no update on irrelevant information)**

Quick advertisement: many more interesting non-quantum examples



Adam Elga (Princeton)

- **Elga's Principle of Indifference:**
Similar centered worlds deserve equal credence.
- **Principle CP (probabilistic consistency)**
- **NUI (no update on irrelevant information)**

Theorem: These three principles cannot hold jointly.

Also, ELGA and NUI imply **Restriction A** for $n=2$ "classically".

Overview

1. Preparation: on **private experiments**

2. Limitations of the third-person view: **Restriction A**

3. A physicalist physician: why Restriction A for $n=1$ is bad

4. An overview on **Algorithmic Idealism**

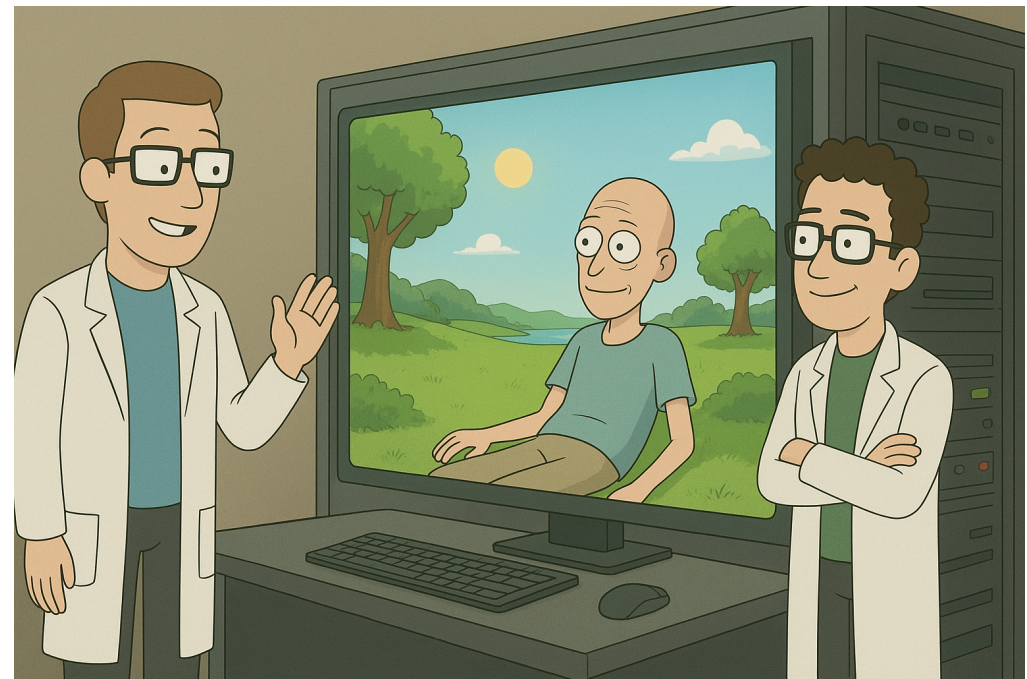
Overview

1. Preparation: on **private experiments**
2. Limitations of the third-person view: **Restriction A**
3. A physicalist physician: why Restriction A for $n=1$ is bad
4. An overview on **Algorithmic Idealism**

A physicalist physician: why **Restriction A** for $n=1$ is bad

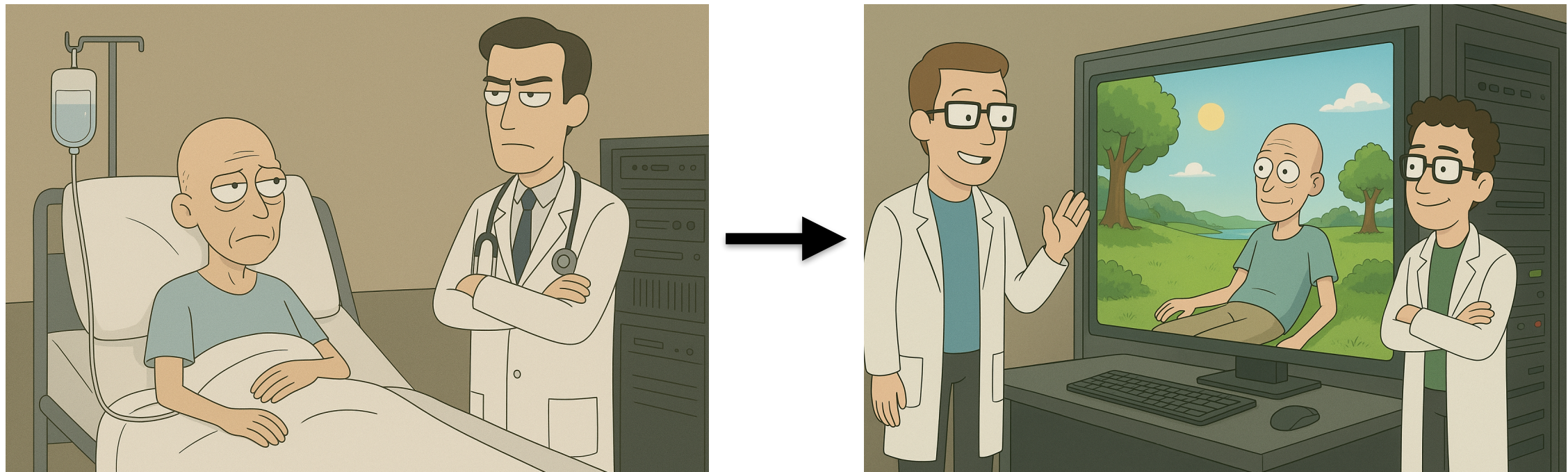
A physicalist physician: why **Restriction A** for $n=1$ is bad

In 2048, you are terminally ill, but you've signed up with a company that will simulate you on a computer after your passing (eliminating the original).



A physicalist physician: why **Restriction A** for $n=1$ is bad

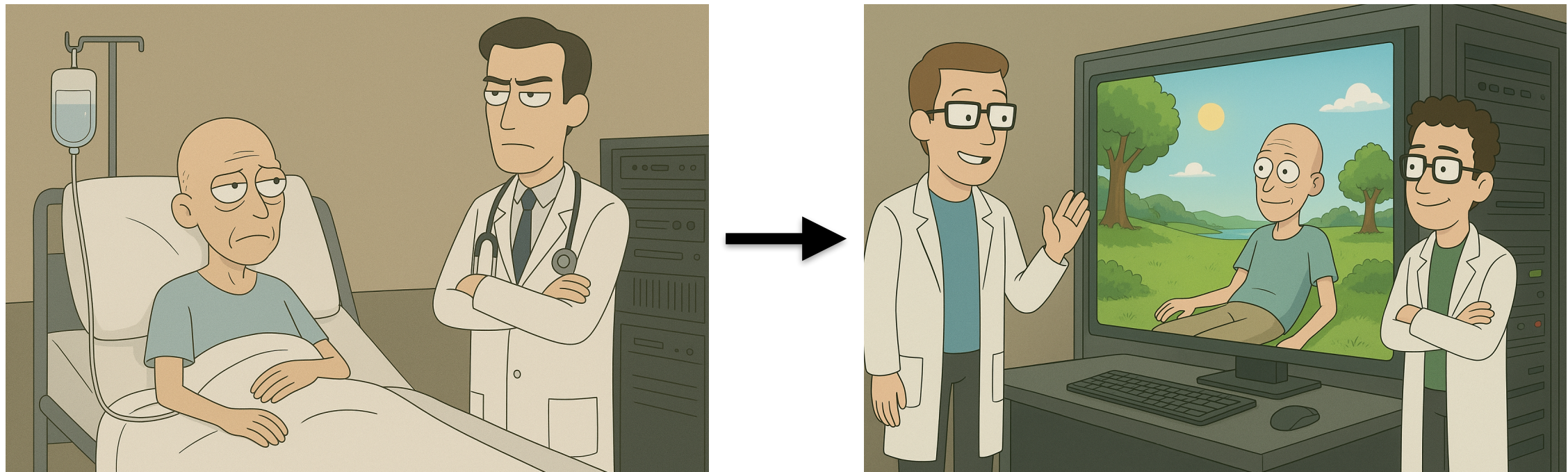
In 2048, you are terminally ill, but you've signed up with a company that will simulate you on a computer after your passing (eliminating the original).



You: Doctor, I'm so scared! I know this is not your speciality, but...
I need to know! Will I really wake up in the computer simulation?

A physicalist physician: why **Restriction A** for $n=1$ is bad

In 2048, you are terminally ill, but you've signed up with a company that will simulate you on a computer after your passing (eliminating the original).



You: Doctor, I'm so scared! I know this is not your speciality, but...
I need to know! Will I really wake up in the computer simulation?

Doctor: Hahaha, you fool! You are asking a non-question! All there is to say is that there is a human being here now, and a computer running a simulation of that thing later. This is all there is to know about the facts of the world. What is the proposition that you are even uncertain *about*?

The lesson I take from this

- Restriction A for $n=1$ agents indicates a **gap in our current understanding** of reality: we generally cannot predict our private futures, even in some cases of great relevance. This needs to be overcome.

The lesson I take from this

- Restriction A for $n=1$ agents indicates a **gap in our current understanding** of reality: we generally cannot predict our private futures, even in some cases of great relevance. This needs to be overcome.

Q: But “*what will I observe?*” may be personally meaningful but still be fundamentally vague or ill-defined. Maybe were asking too much (from physics+) — and *fundamentally*, nothing can be said about it?

The lesson I take from this

- Restriction A for $n=1$ agents indicates a **gap in our current understanding** of reality: we generally cannot predict our private futures, even in some cases of great relevance. This needs to be overcome.

Q: But “*what will I observe?*” may be personally meaningful but still be fundamentally vague or ill-defined. Maybe were asking too much (from physics+) — and *fundamentally*, nothing can be said about it?

A: Everybody can do such a “private experiment”, and Nature will respond. This makes it implausible that nothing can be said about the expected outcome beforehand. Furthermore, the claim that “nothing can be said” does, paradoxically, *already* say something (e.g. a claim of randomness...)

The lesson I take from this

- Restriction A for $n=1$ agents indicates a **gap in our current understanding** of reality: we generally cannot predict our private futures, even in some cases of great relevance. This needs to be overcome.

Q: But “*what will I observe?*” may be personally meaningful but still be fundamentally vague or ill-defined. Maybe were asking too much (from physics+) — and *fundamentally*, nothing can be said about it?

A: Everybody can do such a “private experiment”, and Nature will respond. This makes it implausible that nothing can be said about the expected outcome beforehand. Furthermore, the claim that “nothing can be said” does, paradoxically, *already* say something (e.g. a claim of randomness...)

- Restriction A for $n \geq 2$ agents won't go away. It tells us that we cannot in general deduce the future observations of agents from properties of a (joint) “world” into which they would be embedded.
To “close the gap”, we need to begin with the first-person perspective.

This motivates Algorithmic Idealism

Suggestion: We can see this as motivation to construct “idealist” theories, predicting private chances of future observations, and restricting to our usual (meta)physics in the special case of “public” experiments.

This motivates Algorithmic Idealism

Suggestion: We can see this as motivation to construct “idealist” theories, predicting private chances of future observations, and restricting to our usual (meta)physics in the special case of “public” experiments.

Public experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that can be intersubjectively observed.

Private experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that can ~~(only?)~~ be privately observed.

This motivates Algorithmic Idealism

Suggestion: We can see this as motivation to construct “idealist” theories, predicting private chances of future observations, and restricting to our usual (meta)physics in the special case of “public” experiments.

Public experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that can be intersubjectively observed.

Private experiments:

(Perhaps probabilistic) predictions for the outcomes of experiments that can ~~(only?)~~ be privately observed.

Algorithmic Idealism does exactly that.

It intends to answer the question "What will I see next?" in **all** situations -- exotic or not. In standard "public" situations, it provably agrees with physics. But it also removes n=1 Restriction A **and answers the patient.**

Overview

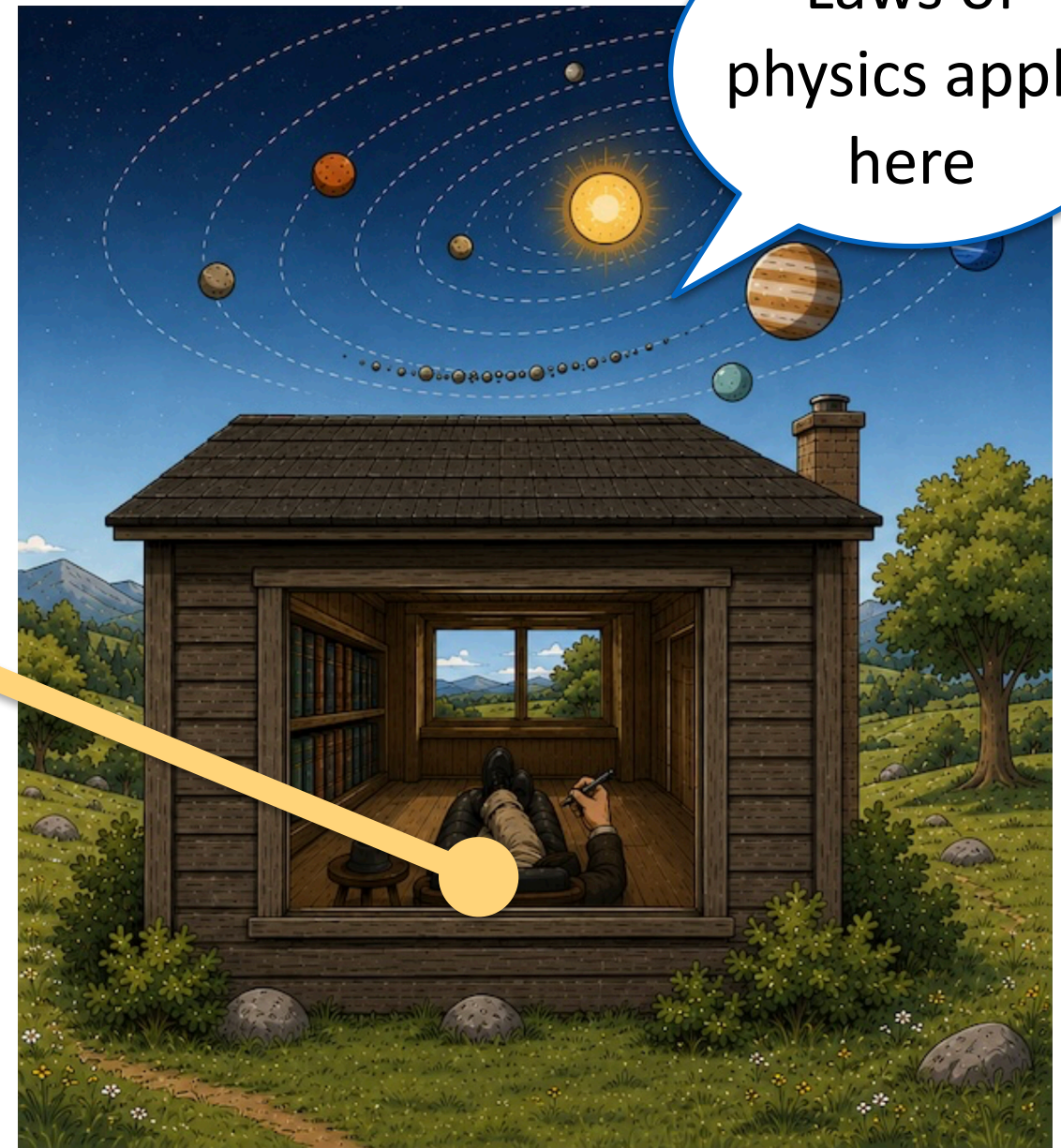
1. Preparation: on **private experiments**
2. Limitations of the third-person view: **Restriction A**
3. A physicalist physician: why Restriction A for $n=1$ is bad
4. An overview on **Algorithmic Idealism**

Overview

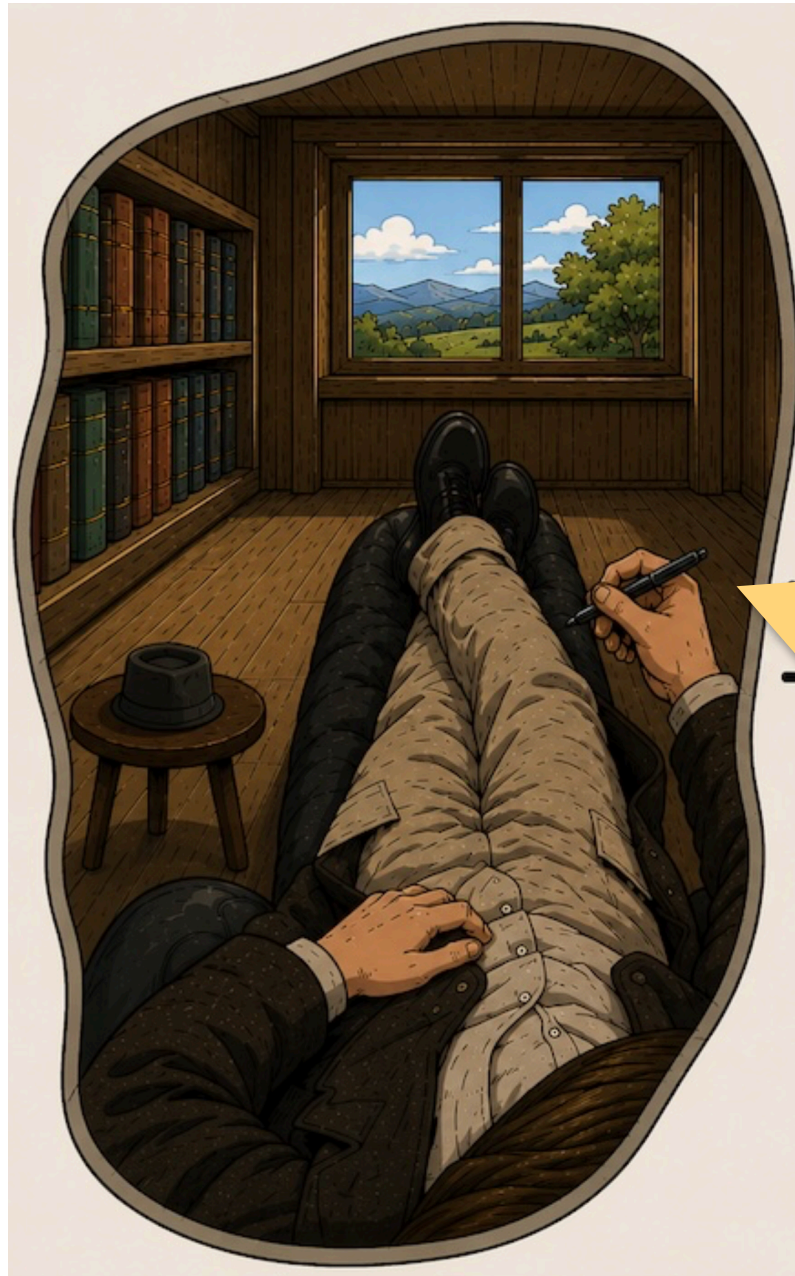
1. Preparation: on **private experiments**
2. Limitations of the third-person view: **Restriction A**
3. A physicalist physician: why Restriction A for $n=1$ is bad
4. An overview on **Algorithmic Idealism**

The standard view

The standard view

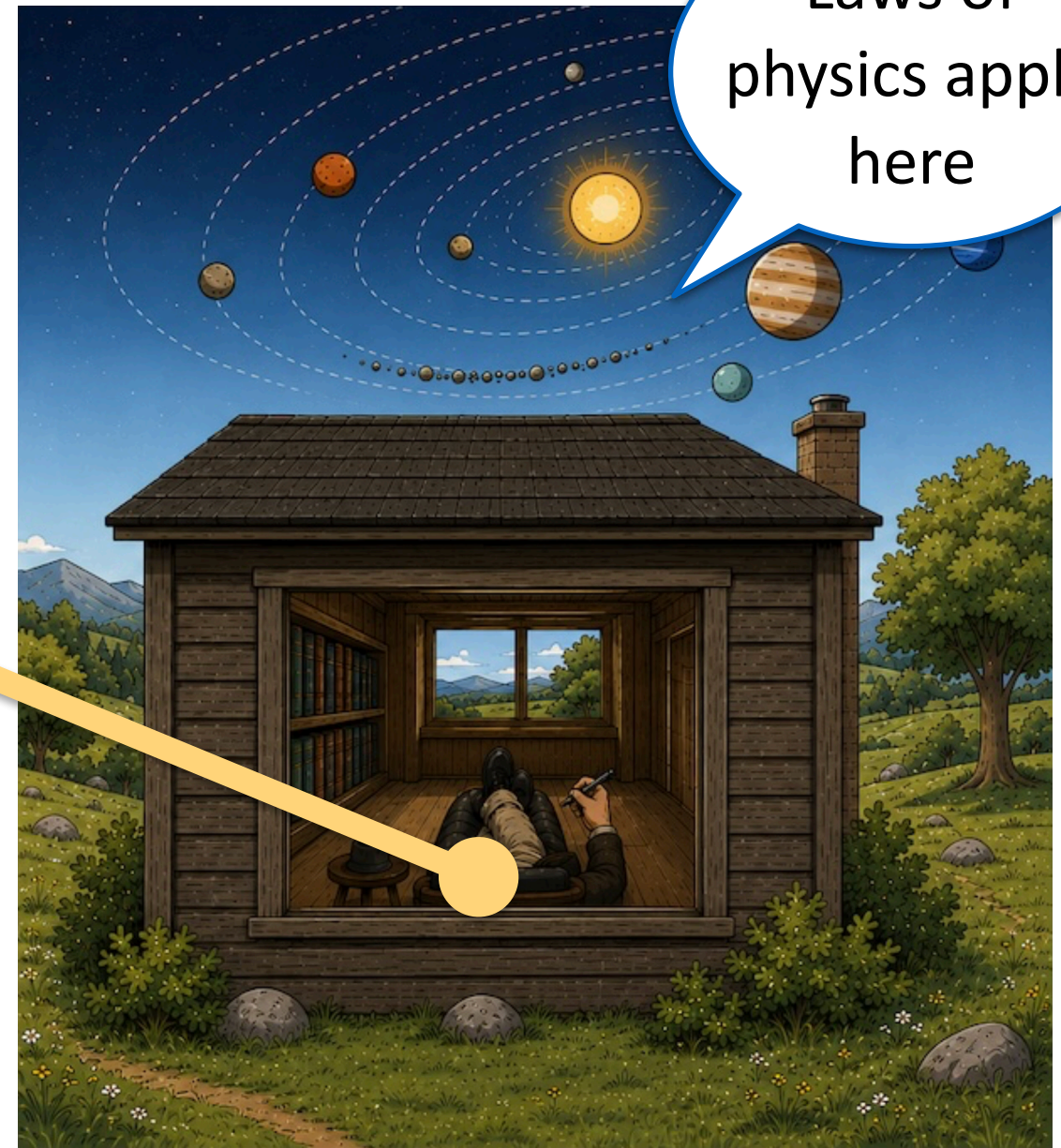


The standard view



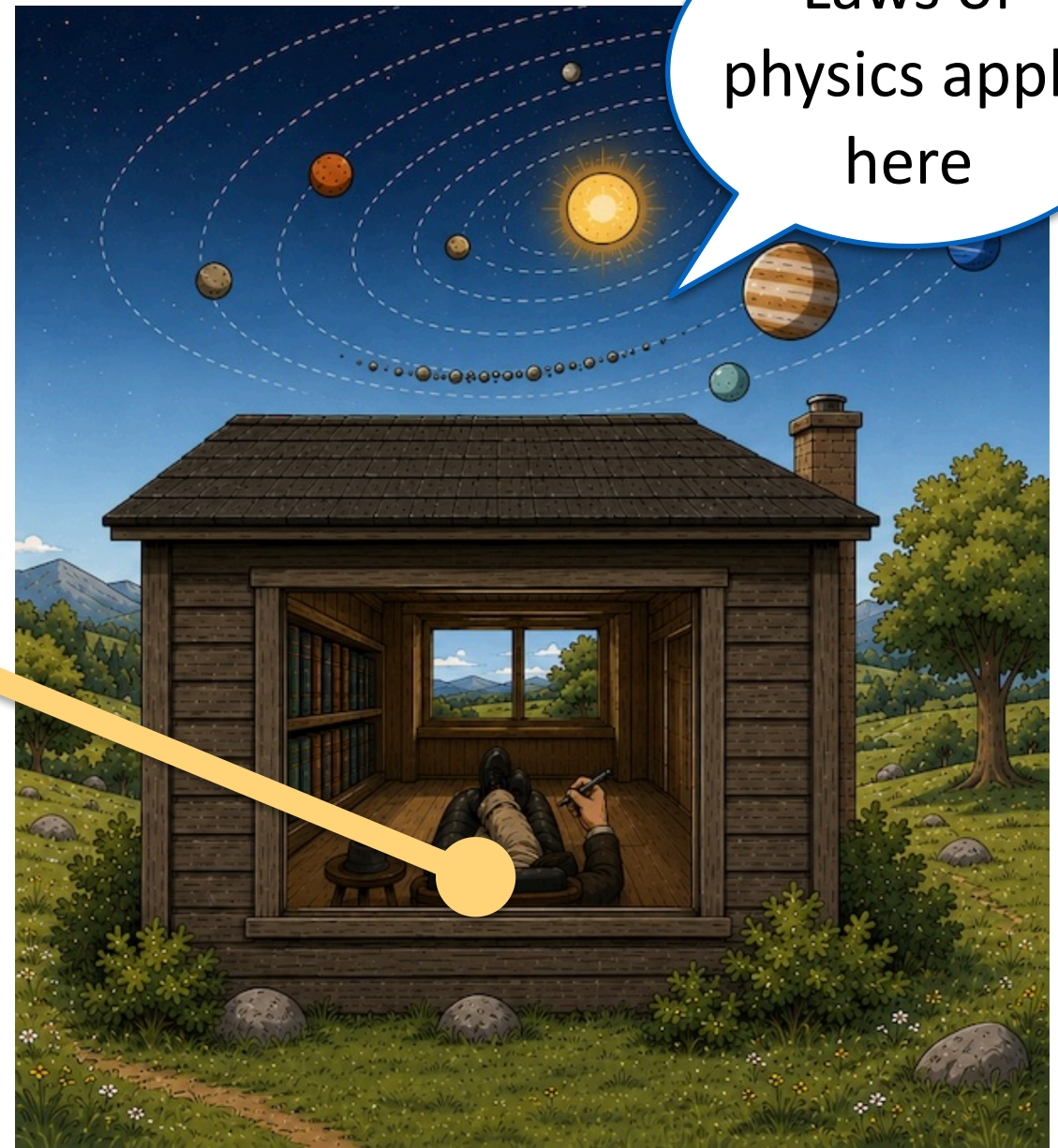
“**self state**” (what “I am now”,
observations+memory+...)

follows from



state (and evolution)
of the **physical world**

The standard view



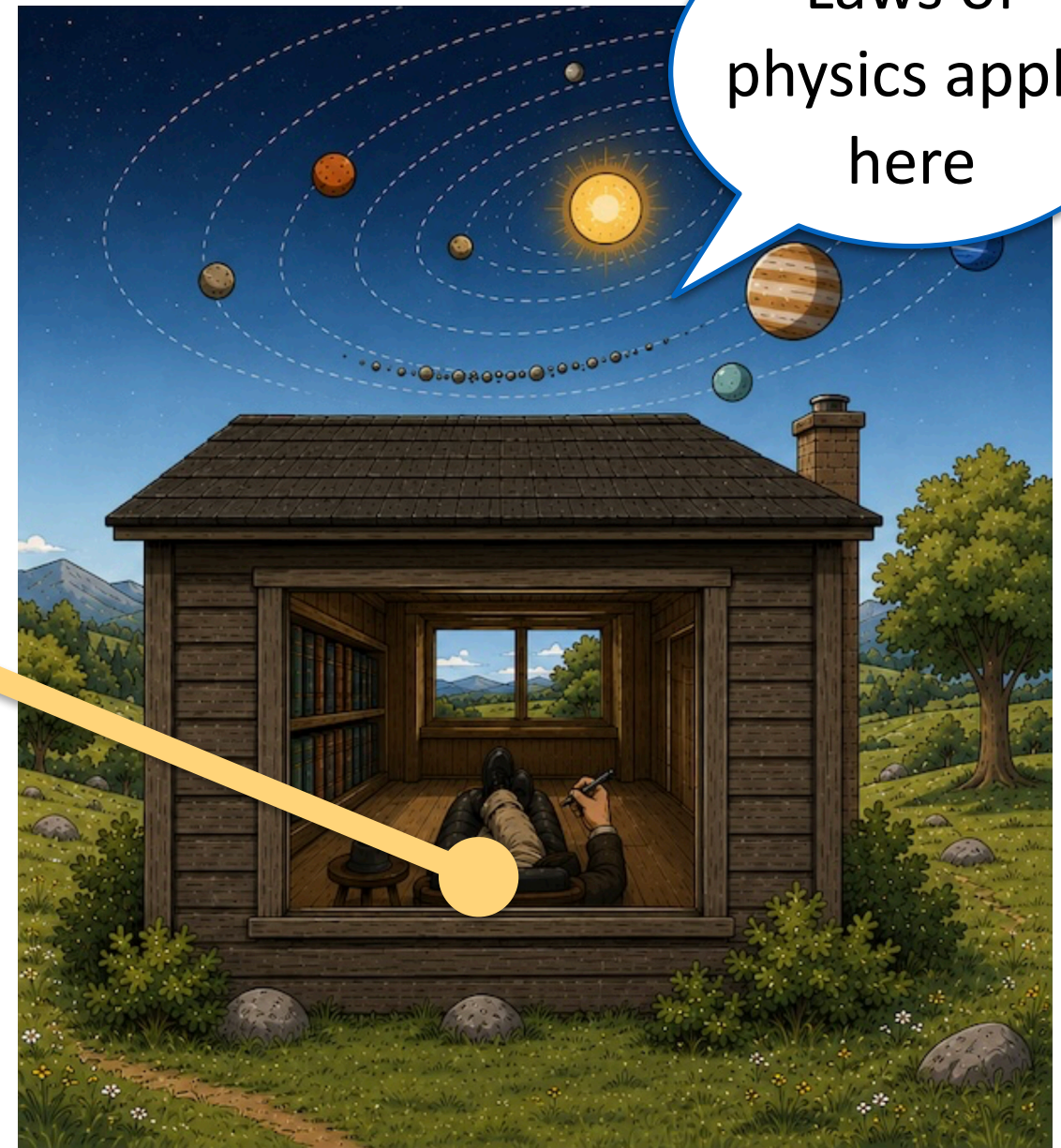
“**self state**” (what “I am now”,
observations+memory+...)

follows from

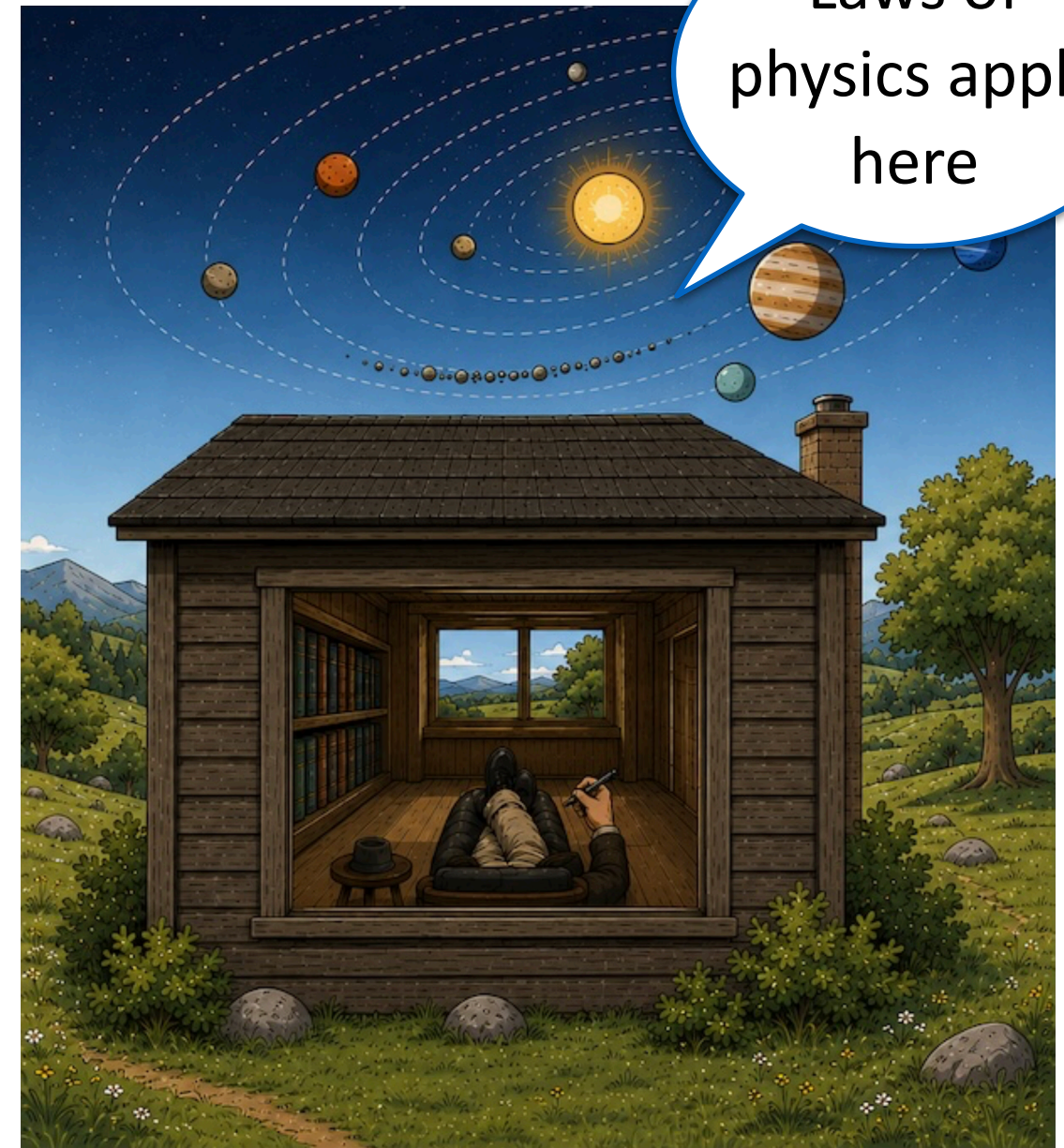
state (and evolution)
of the **physical world**

Problem: Recall **Restriction A** (in particular for $n=1$).
It implies the incompleteness of the standard view.

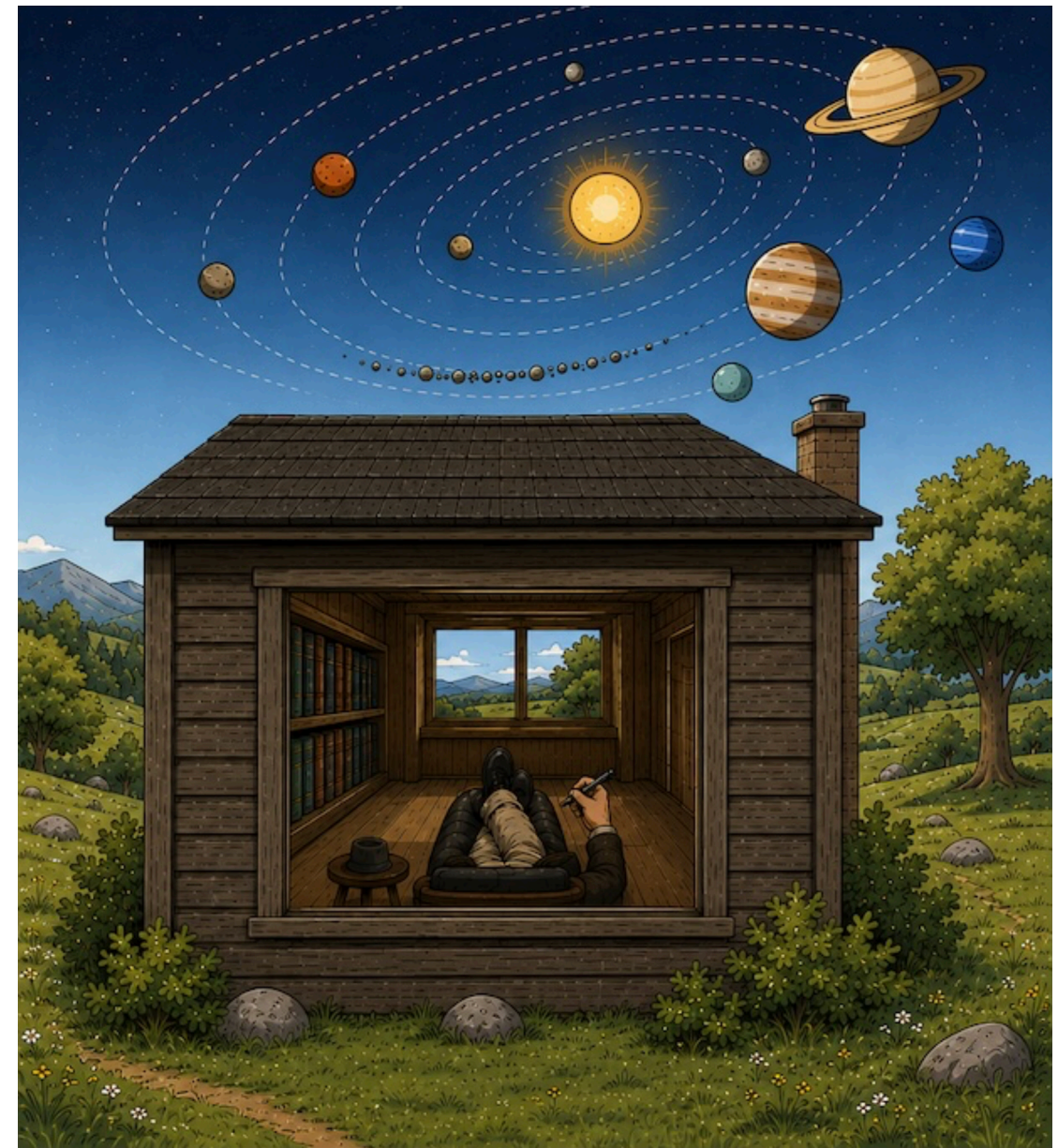
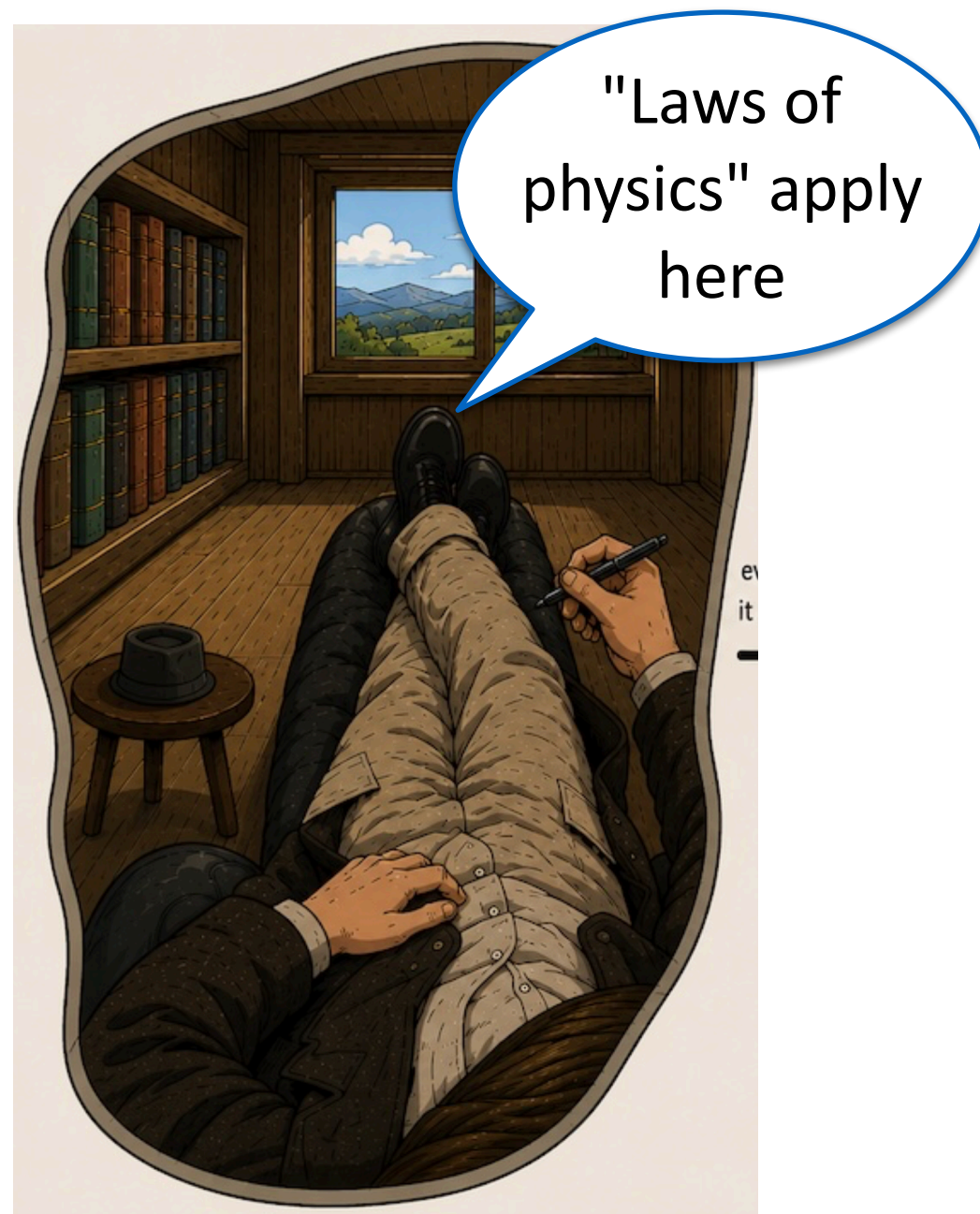
Reversing the standard view



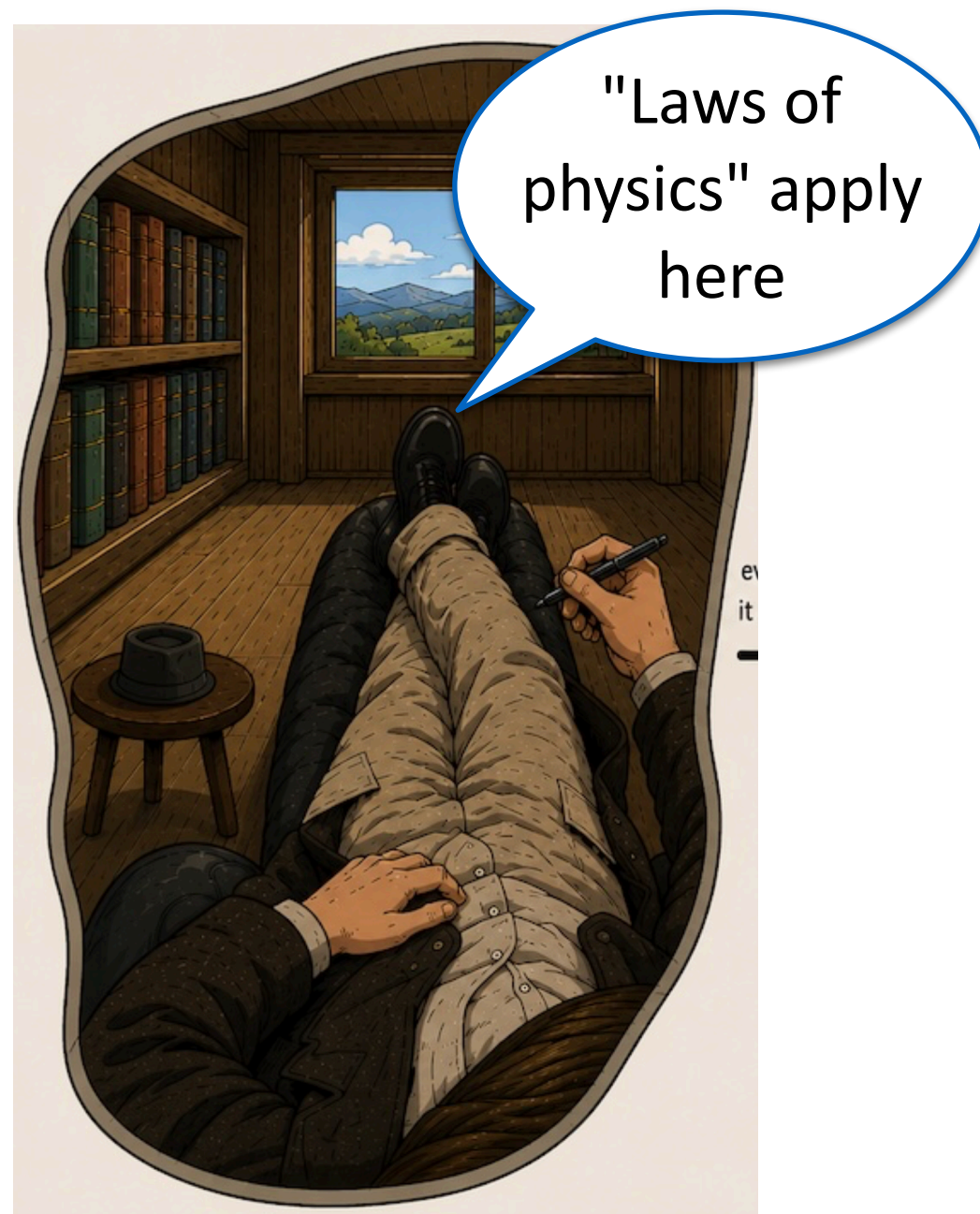
Reversing the standard view



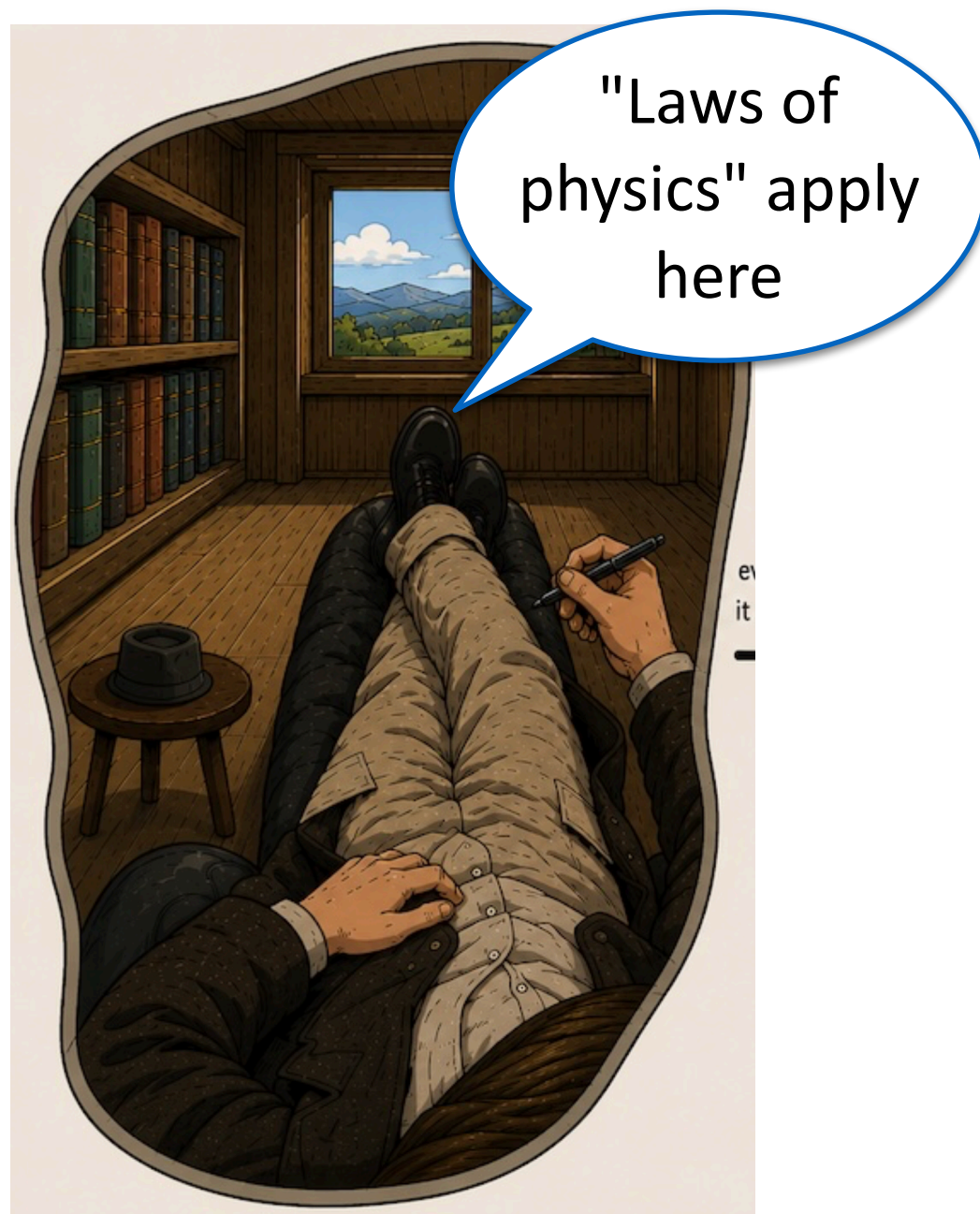
Reversing the standard view



Reversing the standard view



Reversing the standard view



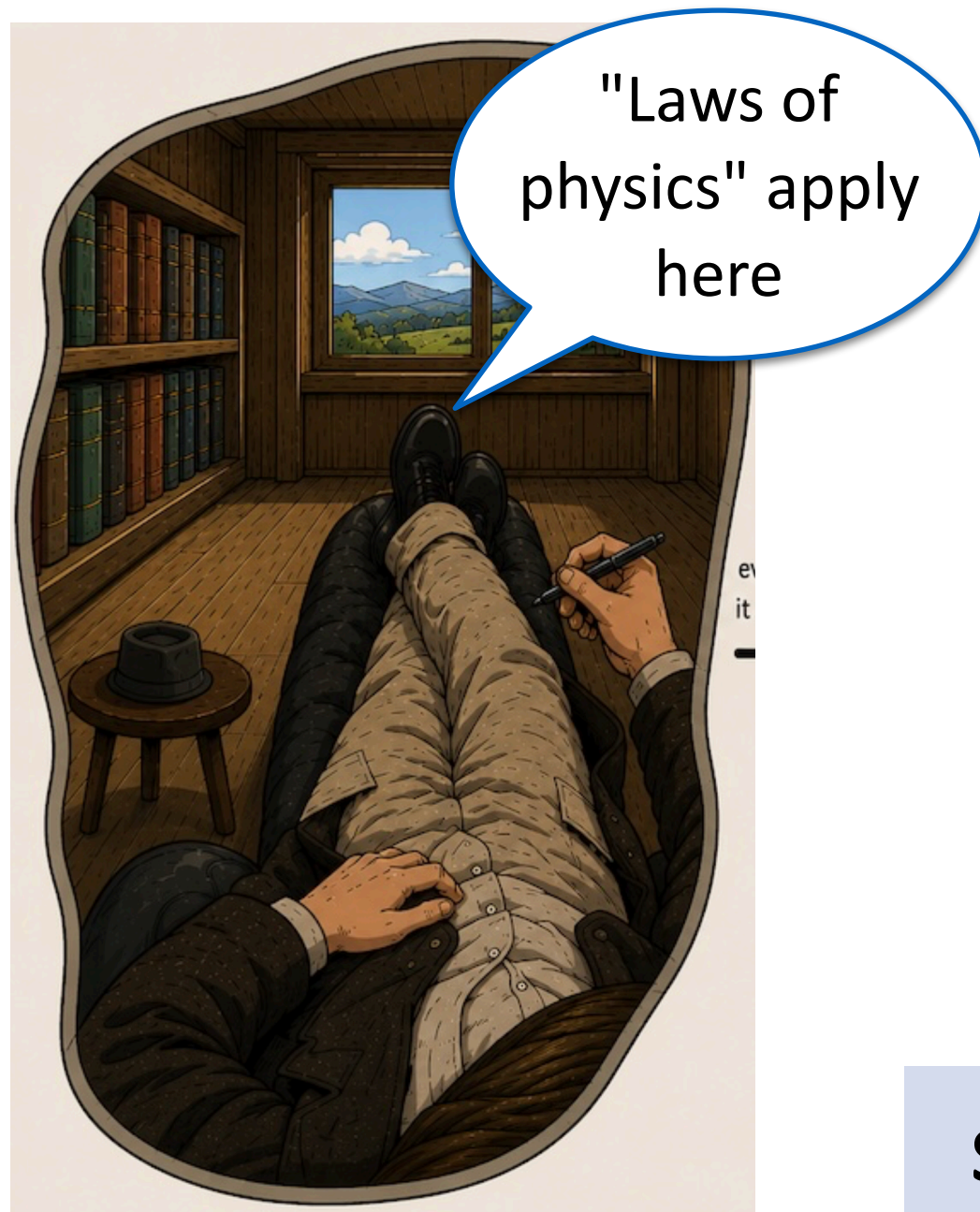
$$\mathbf{P}(y|x)$$

x : self state now

y : self state next

$\mathbf{P}$ is a "universal prior"

Reversing the standard view



$$\mathbf{P}(y|x)$$

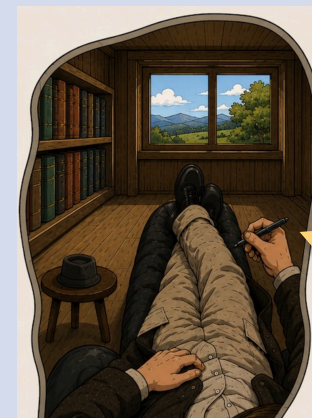
x : self state now

y : self state next

$\mathbf{P}$ is a "universal prior"



Standard view as an approximation:
In the long run, this will *appear as if*

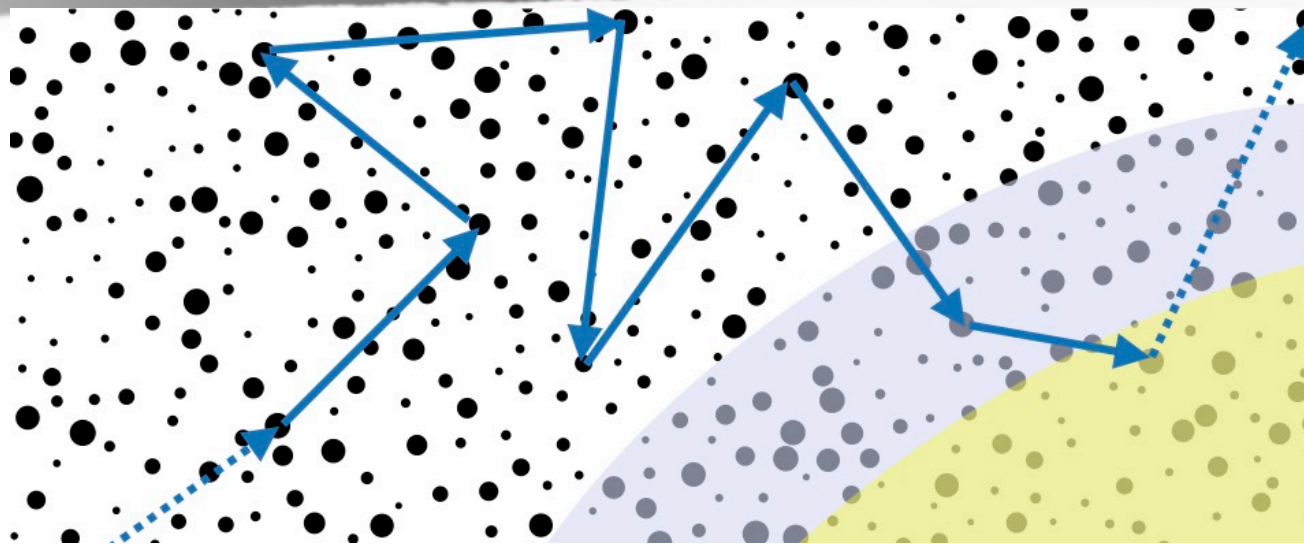


Simple
laws

The postulates of Algorithmic Idealism, and **different models**

The postulates of Algorithmic Idealism, and **different models**

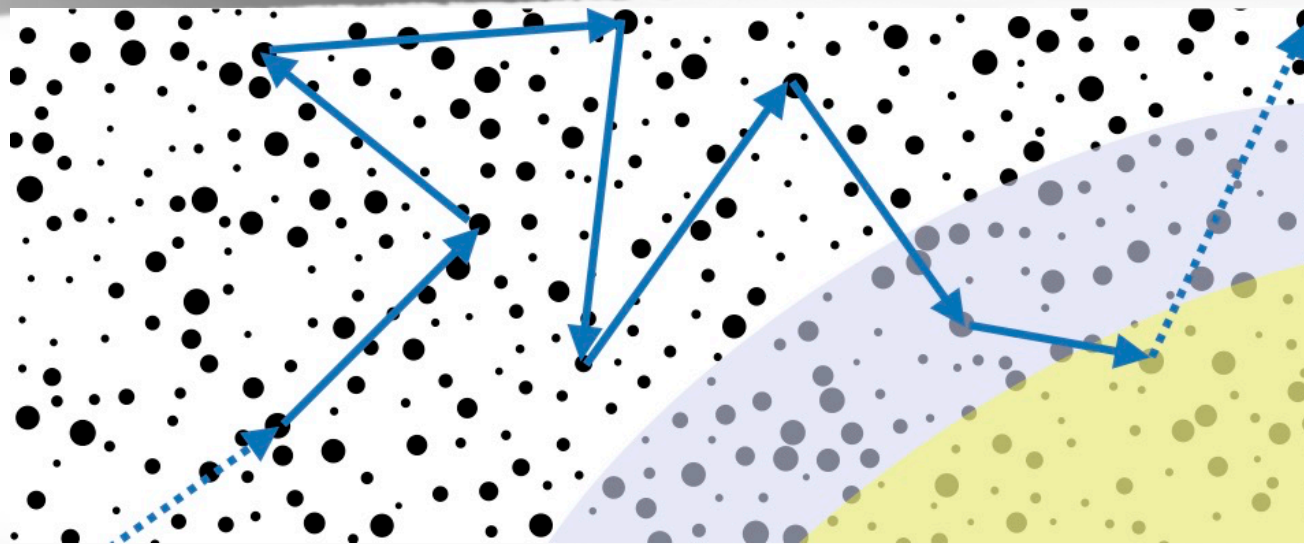
Postulate 1 (Self States). *There is a set S of "self states", the collection of patterns that one can be at any given subjective moment. Everything that is to be said about a given agent at some moment, including all inferences to be drawn about its past or future, is determined by its self state.*



(Most of them are just **abstract patterns** without a natural interpretation of pertaining to "agents" or being related to "consciousness".)

The postulates of Algorithmic Idealism, and **different models**

Postulate 1 (Self States). *There is a set S of "self states", the collection of patterns that one can be at any given subjective moment. Everything that is to be said about a given agent at some moment, including all inferences to be drawn about its past or future, is determined by its self state.*

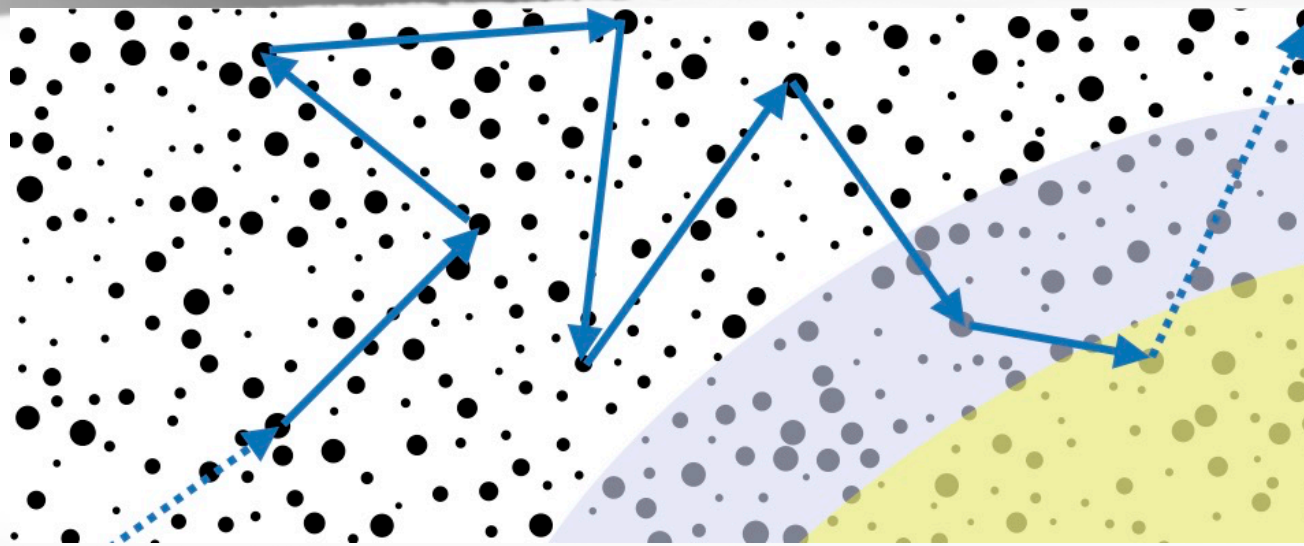


(Most of them are just **abstract patterns** without a natural interpretation of pertaining to "agents" or being related to "consciousness".)

Postulate 2 (State Change). *If you are in self state x now, you will be in another self state y next, and which one this will be manifests itself as a random experiment for you. The likelihood of some y next, given x now, has an objective value, which expresses what you should believe will happen to you next, if you knew your current self state. The mathematical structure that defines its value formalizes a universal method M of induction.*

The postulates of Algorithmic Idealism, and **different models**

Postulate 1 (Self States). *There is a set S of "self states", the collection of patterns that one can be at any given subjective moment. Everything that is to be said about a given agent at some moment, including all inferences to be drawn about its past or future, is determined by its self state.*



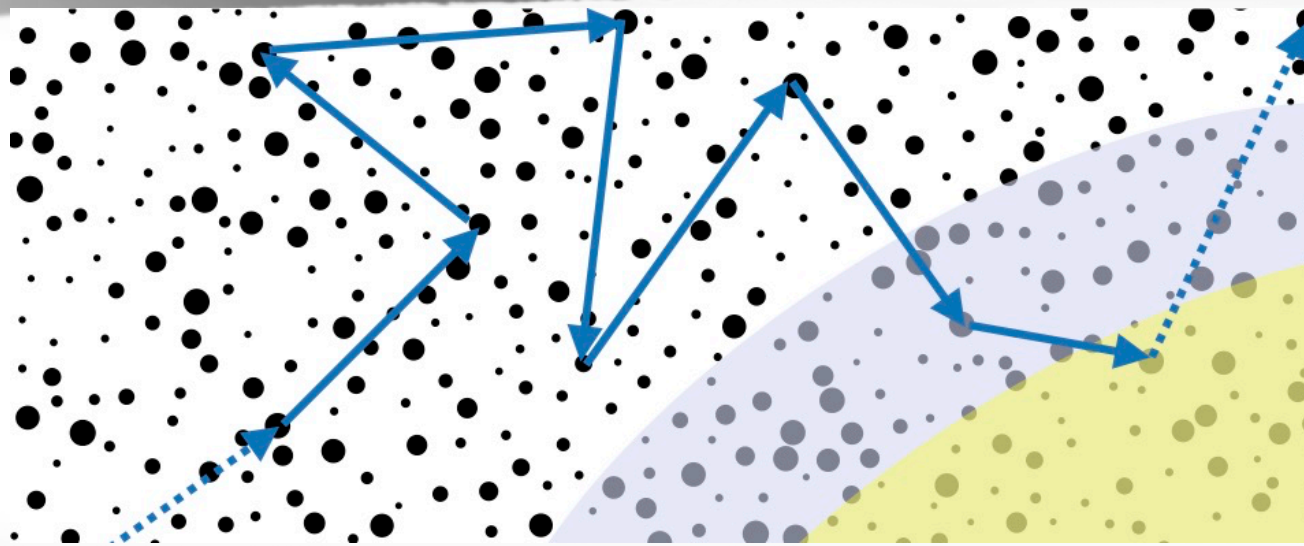
(Most of them are just **abstract patterns** without a natural interpretation of pertaining to "agents" or being related to "consciousness".)

Postulate 2 (State Change). *If you are in self state x now, you will be in another self state y next, and which one this will be manifests itself as a random experiment for you. The likelihood of some y next, given x now, has an objective value, which expresses what you should believe will happen to you next, if you knew your current self state. The mathematical structure that defines its value formalizes a universal method M of induction.*

- **Bit model** (Quantum 4, 301 (2020)): Self states are finite binary strings, and every moment, one learns a new bit that's appended at the end.
Simplification due to the way algorithmic I.T. has so far been formulated.

The postulates of Algorithmic Idealism, and **different models**

Postulate 1 (Self States). *There is a set S of "self states", the collection of patterns that one can be at any given subjective moment. Everything that is to be said about a given agent at some moment, including all inferences to be drawn about its past or future, is determined by its self state.*



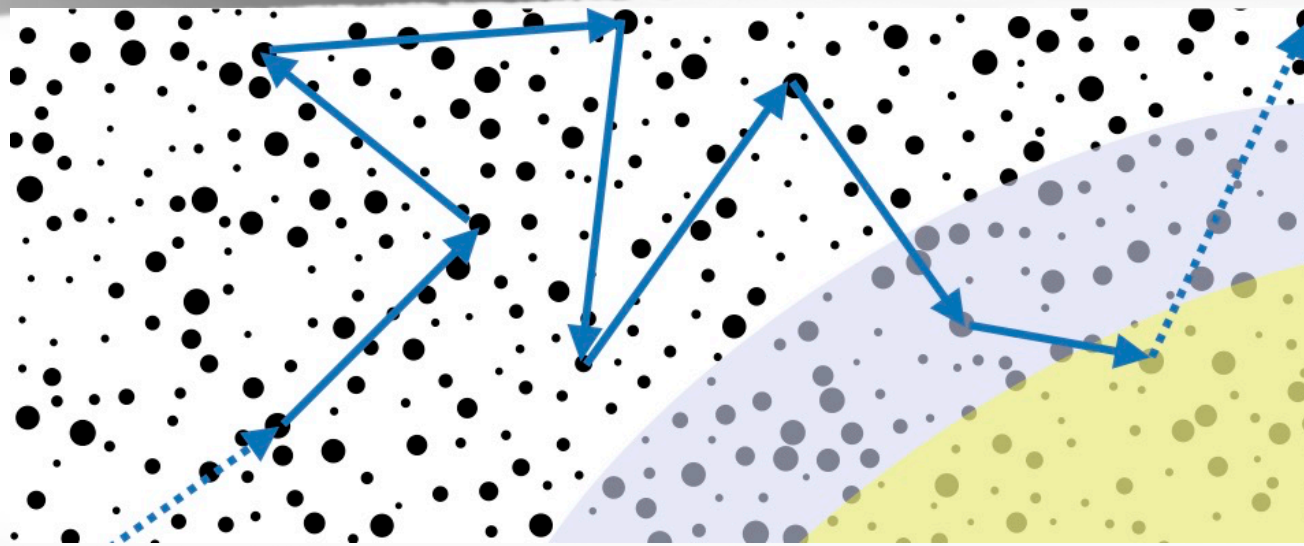
(Most of them are just **abstract patterns** without a natural interpretation of pertaining to "agents" or being related to "consciousness".)

Postulate 2 (State Change). *If you are in self state x now, you will be in another self state y next, and which one this will be manifests itself as a random experiment for you. The likelihood of some y next, given x now, has an objective value, which expresses what you should believe will happen to you next, if you knew your current self state. The mathematical structure that defines its value formalizes a universal method M of induction.*

- **Bit model** (Quantum 4, 301 (2020)): Self states are finite binary strings, and every moment, one learns a new bit that's appended at the end.
Simplification due to the way algorithmic I.T. has so far been formulated.
- **Parfit graph model** (to appear 2026 or 2027). Cannot say much today.

The postulates of Algorithmic Idealism, and **different models**

Postulate 1 (Self States). *There is a set S of "self states", the collection of patterns that one can be at any given subjective moment. Everything that is to be said about a given agent at some moment, including all inferences to be drawn about its past or future, is determined by its self state.*



(Most of them are just **abstract patterns** without a natural interpretation of pertaining to "agents" or being related to "consciousness".)

Postulate 2 (State Change). *If you are in self state x now, you will be in another self state y next, and which one this will be manifests itself as a random experiment for you. The likelihood of some y next, given x now, has an objective value, which expresses what you should believe will happen to you next, if you knew your current self state. The mathematical structure that defines its value formalizes a universal method M of induction.*

- **Bit model** (Quantum 4, 301 (2020)): Self states are finite binary strings, and every moment, one learns a new bit that's appended at the end. **Simplification due to the way algorithmic I.T. has so far been formulated.**
- **Parfit graph model** (to appear 2026 or 2027). Cannot say much today.

Solomonoff induction and the universal prior

Describe **self patterns** as a finite bit strings, growing one bit at a time.

$$x = 01001010001000001111.$$

Solomonoff induction and the universal prior

Describe **self patterns** as a finite bit strings, growing one bit at a time.

$x = 010010100010000011111.$

Q: But I'm not a bunch of bits! So **how** do I **encode** my pattern into bits?

A: It doesn't matter — the resulting theory is **independent of the choice of encoding**, similarly as GR is independent of choice of coordinates.

Solomonoff induction and the universal prior

Describe **self patterns** as a finite bit strings, growing one bit at a time.

$x = 010010100010000011111.$

Q: But I'm not a bunch of bits! So **how** do I **encode** my pattern into bits?

A: It doesn't matter — the resulting theory is **independent of the choice of encoding**, similarly as GR is independent of choice of coordinates.



Ray J. Solomonoff (1964)

Measure: $\mu(b|x) = \frac{\mu(xb)}{\mu(x)}$

Prob. that next bit is b if now in state x .

Solomonoff induction and the universal prior

Describe **self patterns** as a finite bit strings, growing one bit at a time.

$x = 010010100010000011111.$

Q: But I'm not a bunch of bits! So **how** do I **encode** my pattern into bits?

A: It doesn't matter — the resulting theory is **independent of the choice of encoding**, similarly as GR is independent of choice of coordinates.



Ray J. Solomonoff (1964)

Measure: $\mu(b|x) = \frac{\mu(xb)}{\mu(x)}$

Prob. that next bit is b if now in state x .

$$\mu(0|x) + \mu(1|x) = 1.$$

Solomonoff induction and the universal prior

Describe **self patterns** as a finite bit strings, growing one bit at a time.

$x = 010010100010000011111.$

Q: But I'm not a bunch of bits! So **how** do I **encode** my pattern into bits?

A: It doesn't matter — the resulting theory is **independent of the choice of encoding**, similarly as GR is independent of choice of coordinates.



Ray J. Solomonoff (1964)

Measure: $\mu(b|x) = \frac{\mu(xb)}{\mu(x)}$

Prob. that next bit is b if now in state x .

$$\mu(0|x) + \mu(1|x) = 1.$$

Semimeasure: $\mu(0|x) + \mu(1|x) \leq 1.$

Solomonoff induction and the universal prior

Describe **self patterns** as a finite bit strings, growing one bit at a time.

$x = 010010100010000011111$.

Q: But I'm not a bunch of bits! So **how** do I **encode** my pattern into bits?

A: It doesn't matter — the resulting theory is **independent of the choice of encoding**, similarly as GR is independent of choice of coordinates.



Measure: $\mu(b|x) = \frac{\mu(xb)}{\mu(x)}$

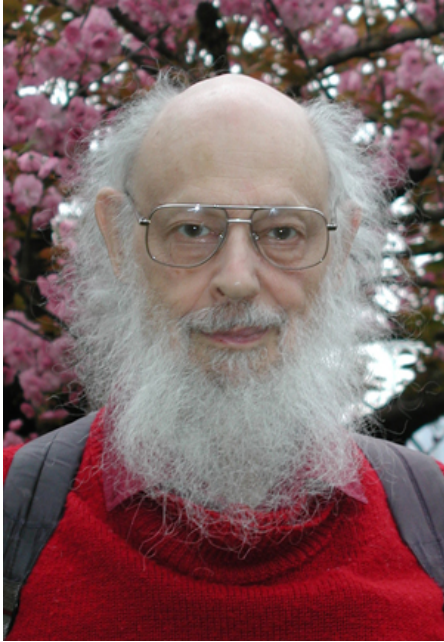
Prob. that next bit is b if now in state x .

$$\mu(0|x) + \mu(1|x) = 1.$$

Semimeasure: $\mu(0|x) + \mu(1|x) \leq 1.$

Enumerable semimeasure: there exists an algorithm that, on input x and n , computes an approx. $\mu_n(x)$ with $\lim_{n \rightarrow \infty} \mu_n(x) = \mu(x)$ and $\mu_1 \leq \mu_2 \leq \dots$

Solomonoff induction and the universal prior



Universal enumerable semimeasure μ :

Solomonoff induction and the universal prior



Universal enumerable semimeasure μ :

For every enumerable semimeasure ν there is a constant $c_\nu > 0$ such that $\mu(x) \geq c_\nu \cdot \nu(x)$ for all x .

Basically, a mixture of all enumerable semimeasures.

Solomonoff induction and the universal prior



Universal enumerable semimeasure μ :

For every enumerable semimeasure ν there is a constant $c_\nu > 0$ such that $\mu(x) \geq c_\nu \cdot \nu(x)$ for all x .

Basically, a mixture of all enumerable semimeasures.

normalize it $\longrightarrow$ **universal probability P .**

Solomonoff induction and the universal prior



Universal enumerable semimeasure μ :

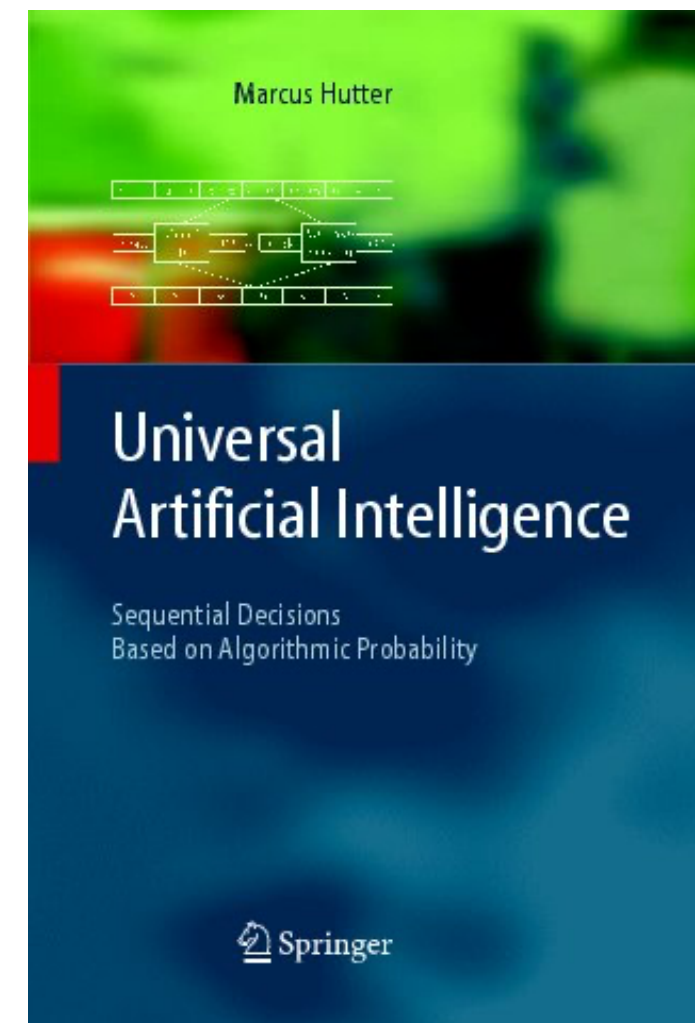
For every enumerable semimeasure ν there is a constant $c_\nu > 0$ such that $\mu(x) \geq c_\nu \cdot \nu(x)$ for all x .

Basically, a mixture of all enumerable semimeasures.

normalize it $\longrightarrow$ **universal probability P .**

Application elsewhere (not in my approach):

- Gives higher probability to simpler bit strings (i.e. generated by shorter programs). **Occam's razor.**
- Uncomputable, but in principle useful for **induction** $\longrightarrow$ "Universal Artificial Intelligence"
- **Solomonoff induction:** yields provably correct predictions asymptotically (quickly) in all computable environments.



Algorithmic idealism's **bit model**

Algorithmic idealism's **bit model**

At every (subjective) moment, “I” am some self-state x , and a subjective moment later, I will be some self-state xy with universal probability $\mathbf{P}(y|x)$.

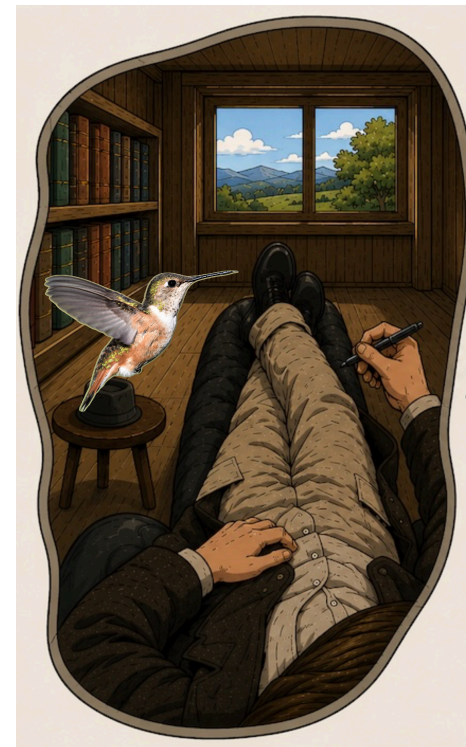
Algorithmic idealism's **bit model**

At every (subjective) moment, “I” am some self-state x , and a subjective moment later, I will be some self-state xy with universal probability $\mathbf{P}(y|x)$.



$$x = 1001$$

$$\mathbf{P}(y|x)$$



$$x' = 1001111101 = xy$$

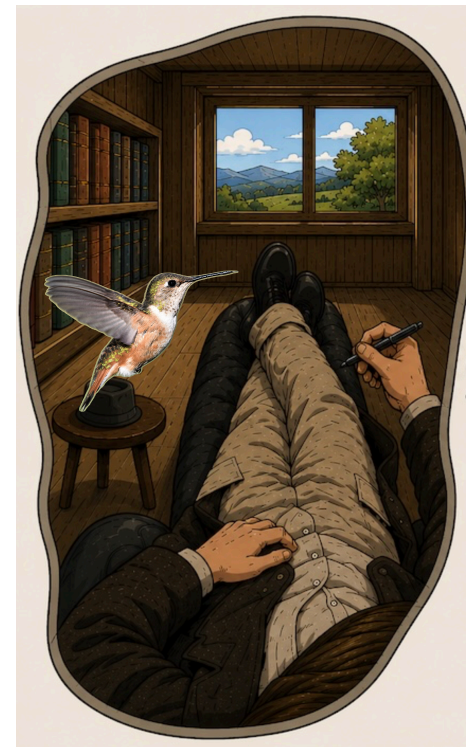
Algorithmic idealism's **bit model**

At every (subjective) moment, “I” am some self-state x , and a subjective moment later, I will be some self-state xy with universal probability $\mathbf{P}(y|x)$.



$$x = 1001$$

$$\mathbf{P}(y|x)$$



$$x' = 1001111101 = xy$$

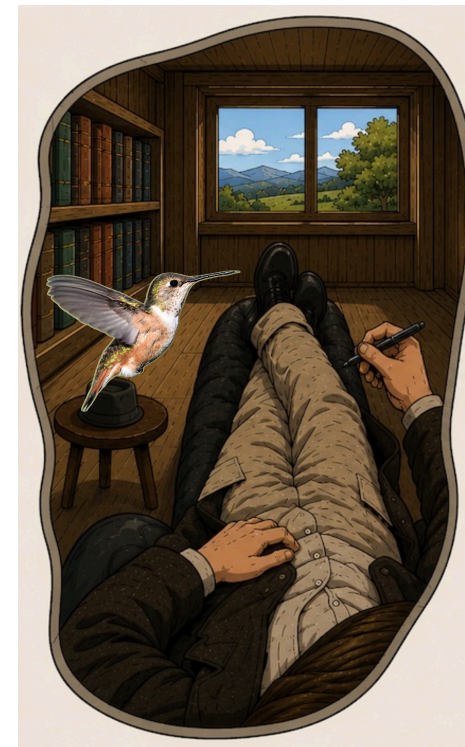
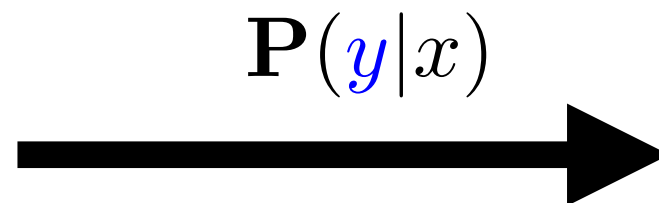
This is a fundamental, objective, private chance that does not arise from any lack of knowledge, or any “external world” in which my self state would be embedded. **“I am an unembedded pattern”.**

Algorithmic idealism's **bit model**

At every (subjective) moment, “I” am some self-state x , and a subjective moment later, I will be some self-state xy with universal probability $\mathbf{P}(y|x)$.



$$x = 1001$$



$$x' = 1001111101 = xy$$

This is a fundamental, objective, private chance that does not arise from any lack of knowledge, or any “external world” in which my self state would be embedded. **“I am an unembedded pattern”**.

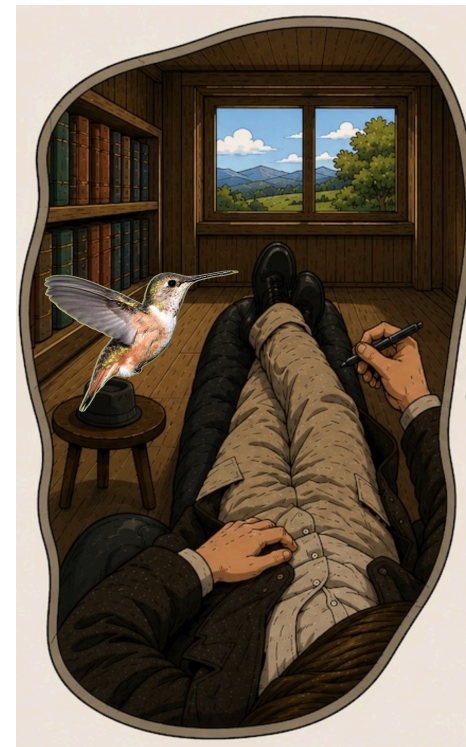
(The bit model is an incomplete simplified model because it does not allow fundamental "forgetting" of information.)

Consistency with standard physics



$$x = 1001$$

$$\mathbf{P}(y|x)$$



$$x' = 10011111101 = xy$$

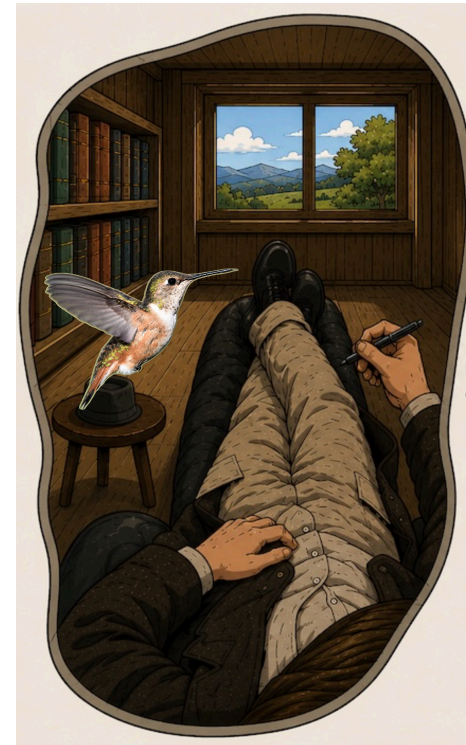
Consistency with standard physics

The standard view would tell us to compute the “physical” probability $\mathbf{P}_{\text{phys}}(\textcolor{blue}{y}|x)$ arising from, say, statistical or quantum mechanics.



$$x = 1001$$

$$\mathbf{P}(\textcolor{blue}{y}|x)$$



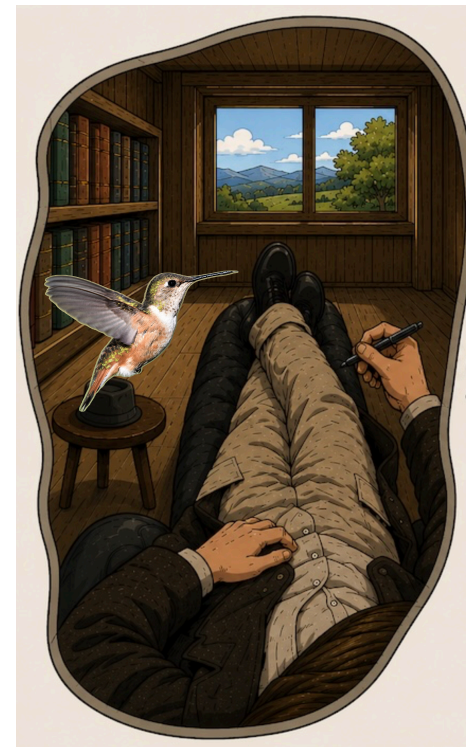
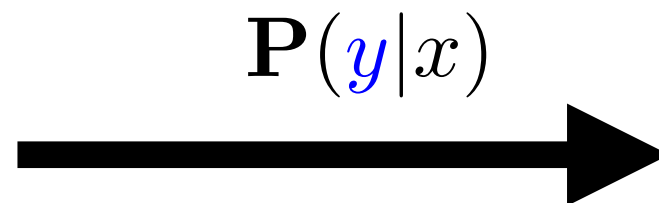
$$x' = 1001\textcolor{blue}{1111101} = x\textcolor{blue}{y}$$

Consistency with standard physics

The standard view would tell us to compute the “physical” probability $\mathbf{P}_{\text{phys}}(\textcolor{blue}{y}|x)$ arising from, say, statistical or quantum mechanics.



$$x = 1001$$



$$x' = 1001\textcolor{blue}{1111101} = x\textcolor{blue}{y}$$

Theorem. In the limit of a large number $n = \ell(x)$ of self state bits,

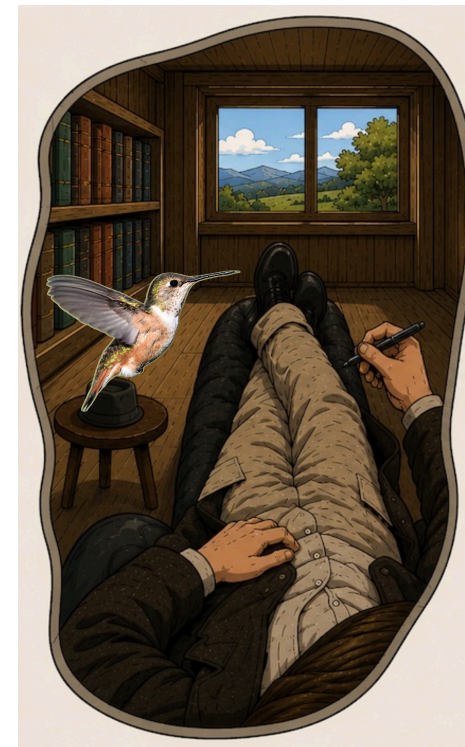
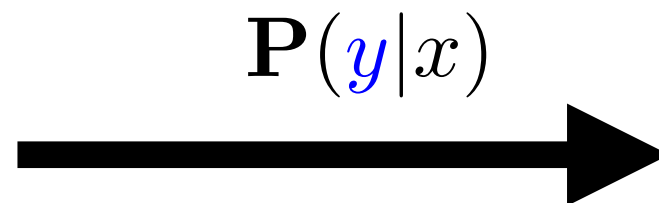
$$|\mathbf{P}(\textcolor{blue}{y}|x) - \mathbf{P}_{\text{phys}}(\textcolor{blue}{y}|x)| \xrightarrow{n \rightarrow \infty} 0.$$

Consistency with standard physics

The standard view would tell us to compute the “physical” probability $\mathbf{P}_{\text{phys}}(\textcolor{blue}{y}|x)$ arising from, say, statistical or quantum mechanics.



$$x = 1001$$



$$x' = 1001\textcolor{blue}{1111101} = x\textcolor{blue}{y}$$

Theorem. In the limit of a large number $n = \ell(x)$ of self state bits,

$$|\mathbf{P}(\textcolor{blue}{y}|x) - \mathbf{P}_{\text{phys}}(\textcolor{blue}{y}|x)| \xrightarrow{n \rightarrow \infty} 0.$$

Proof. Physical versions of the Church-Turing thesis

$\Rightarrow \mathbf{P}_{\text{phys}}$ is in principle computable. Thus, due to Solomonoff's universal induction, convergence above happens with $\mathbf{P}_{\text{phys}}$ -prob. 1.

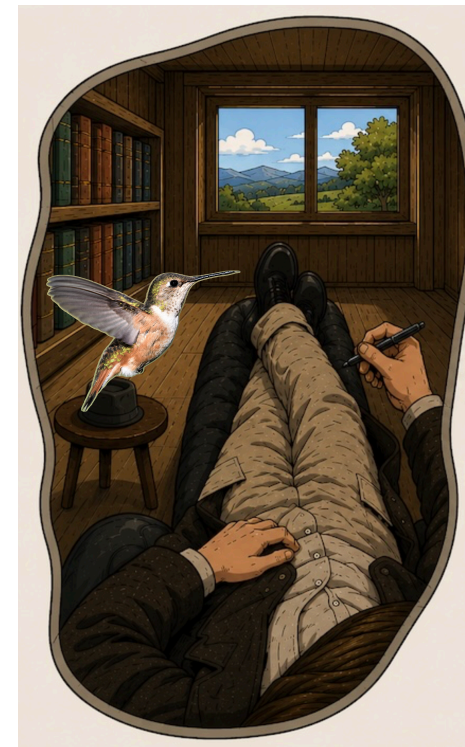
Consistency with standard physics

The standard view would tell us to compute the “physical” probability $\mathbf{P}_{\text{phys}}(\textcolor{blue}{y}|x)$ arising from, say, statistical or quantum mechanics.



$x = 1001$

$\mathbf{P}(\textcolor{blue}{y}|x)$



$x' = 1001\textcolor{blue}{1111101} = x\textcolor{blue}{y}$

Interpretation. If the self state contains enough information on the (for me) relevant aspects of the physical world, then universal probability will “detect” these regularities (**Solomonoff induction**) and assign high probability to the fact that these regularities will remain present.
Hence, physical and universal probabilities will agree in their predictions.

Consistency with standard physics

Consistency with standard physics is good news, but it is **not enough**.

Consistency with standard physics

Consistency with standard physics is good news, but it is **not enough**.

Now that I hold a large amount of information on a (possible) external physical world, universal probability predicts chances that agree with those determined by that world in the future. Fair enough.

But why should I get there in the first place if the universal prior is all there is, and no external world is assumed to begin with?

Consistency with standard physics

Consistency with standard physics is good news, but it is **not enough**.

Now that I hold a large amount of information on a (possible) external physical world, universal probability predicts chances that agree with those determined by that world in the future. Fair enough.

But why should I get there in the first place if the universal prior is all there is, and no external world is assumed to begin with?

As we will now show, universal probability predicts an “external world”.

Consistency with standard physics

Consistency with standard physics is good news, but it is **not enough**.

Now that I hold a large amount of information on a (possible) external physical world, universal probability predicts chances that agree with those determined by that world in the future. Fair enough.

But why should I get there in the first place if the universal prior is all there is, and no external world is assumed to begin with?

As we will now show, universal probability predicts an “external world”.



This does **not** make it a “theory of everything” because it cannot predict most properties of that world.

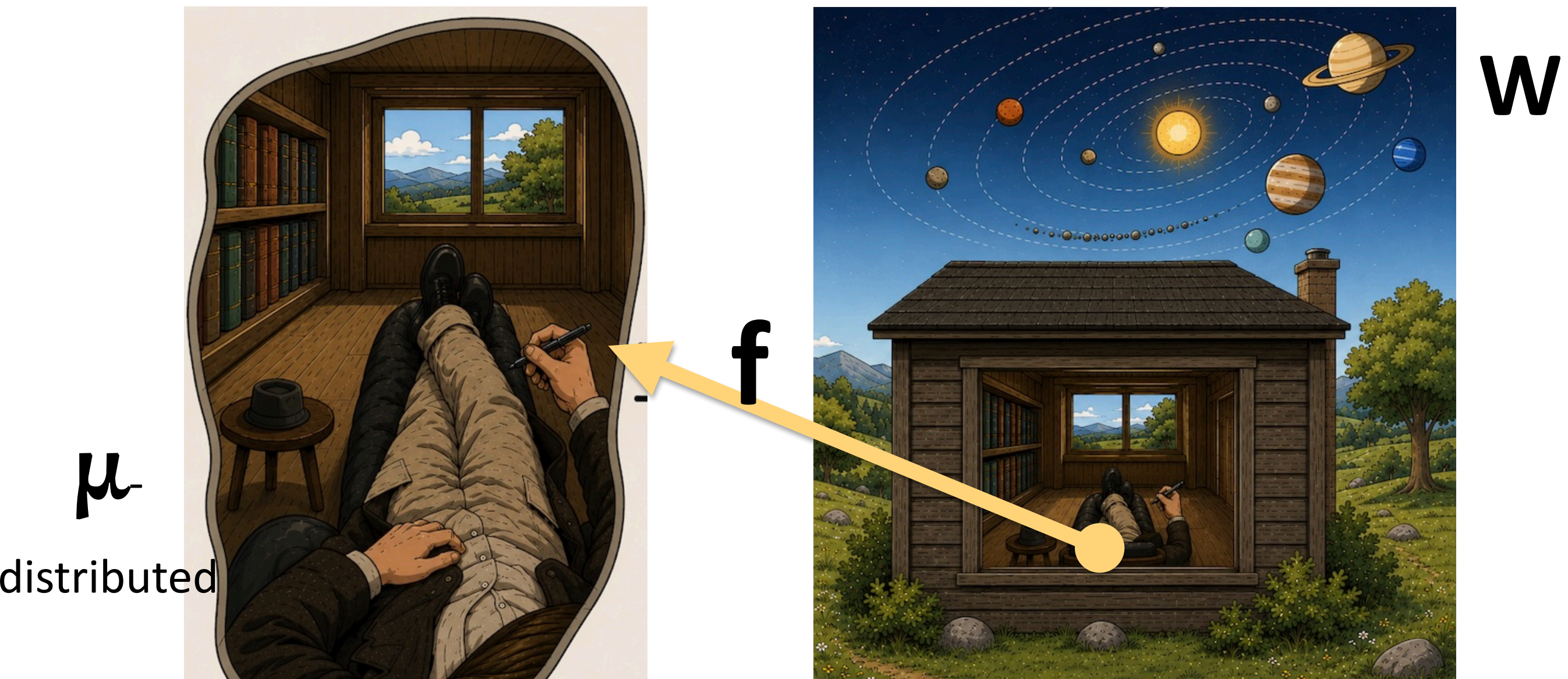
Candidate external worlds

Candidate external worlds

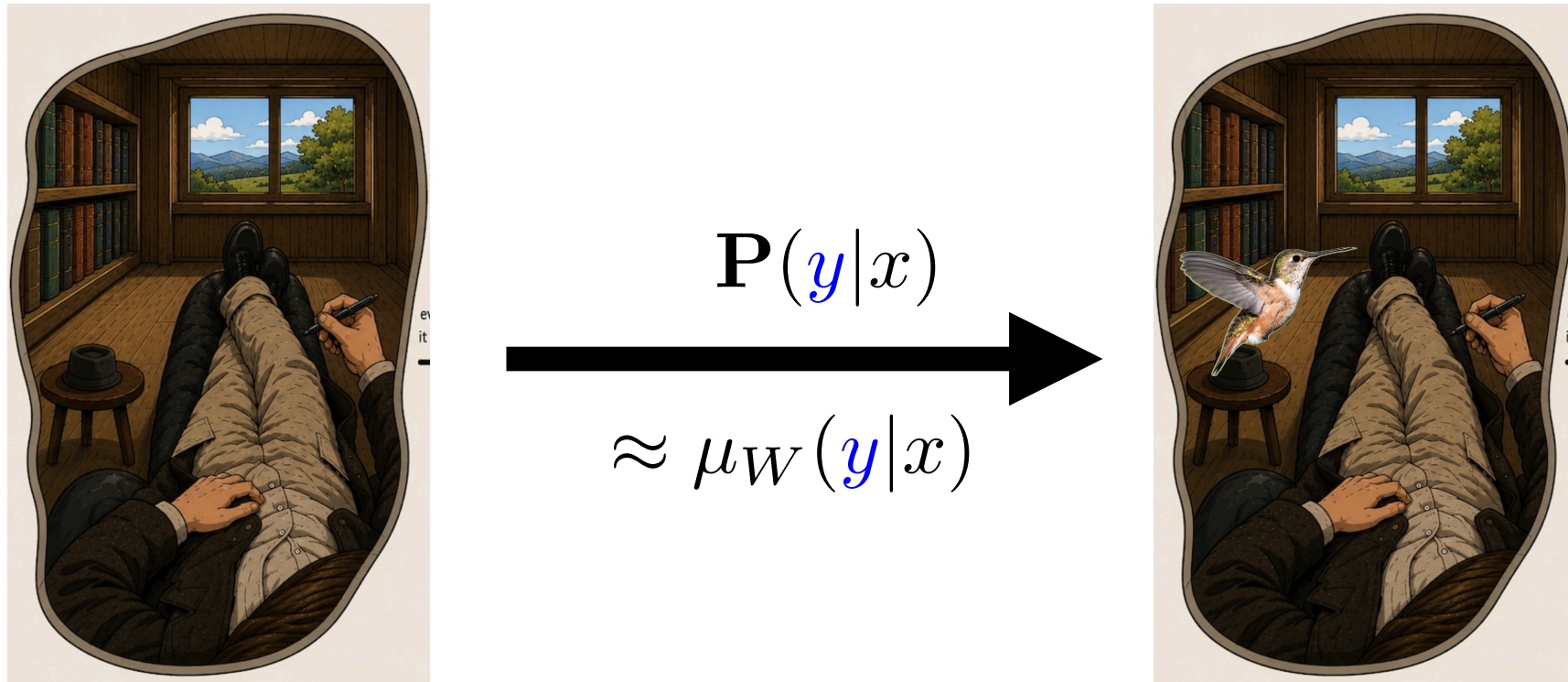
Def.: A **computational ontological model** for μ is a stochastic process (“world” W) that can in principle be run on a probabilistic Turing machine, together with a **computable** bit-string-valued random variable f (intuitively, it “*locates / reads the self state from world W* ”) that yields self states sampled according to μ .

Candidate external worlds

Def.: A **computational ontological model** for μ is a stochastic process (“world” W) that can in principle be run on a probabilistic Turing machine, together with a **computable** bit-string-valued random variable f (intuitively, it “*locates / reads the self state from world W* ”) that yields self states sampled according to μ .



An emergent notion of **external world**

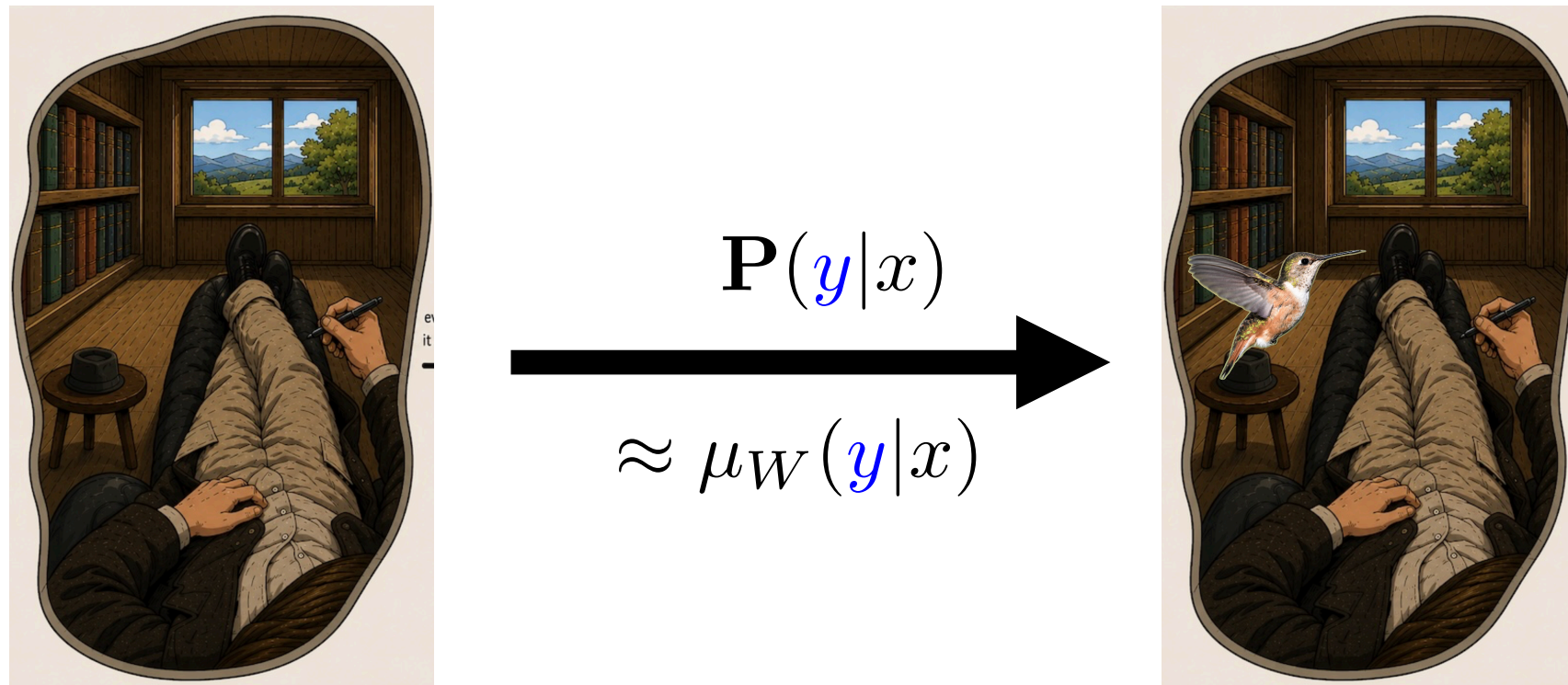


Theorem: Initially, when the self state is the empty string,
there is universal probability $\mathbf{P}$ of at least

$$2^{-K(W)}$$

that ontological model (“world”) W is seen asymptotically,
i.e. that $|\mathbf{P}(y|x_1, \dots, x_n) - \mu_W(y|x_1, \dots, x_n)| \longrightarrow 0$.

An emergent notion of **external world**



Theorem: Initially, when the self state is the empty string,
 there is universal probability $\mathbf{P}$ of at least
 $2^{-\boxed{\mathbf{K}(W)}}$ ← description length of W
 on a universal computer

that ontological model (“world”) W is seen asymptotically,
 i.e. that $|\mathbf{P}(y|x_1, \dots, x_n) - \mu_W(y|x_1, \dots, x_n)| \longrightarrow 0$.

actual chances
 according to univ. probability

chances as determined by
 world W .

An emergent notion of **external world**

Properties of this (**probabilistic**) world W :

- $K(W)$ probably small: W has **simple** “laws of nature”.
- **Actual realization** seen by agent typically **complex** (compare: coin toss).
- In particular, μ_W is probabilistically **computable** (recall: $\mathbf{P}$ isn't!)
- Such processes typically start in a **state of low complexity**. "Big bang"?

Broadly consistent with what we observe!

Theorem: Initially, when the self state is the empty string,

there is universal probability $\mathbf{P}$ of at least

$$2^{-K(W)}$$

description length of W
on a universal computer

that ontological model (“world”) W is seen asymptotically,

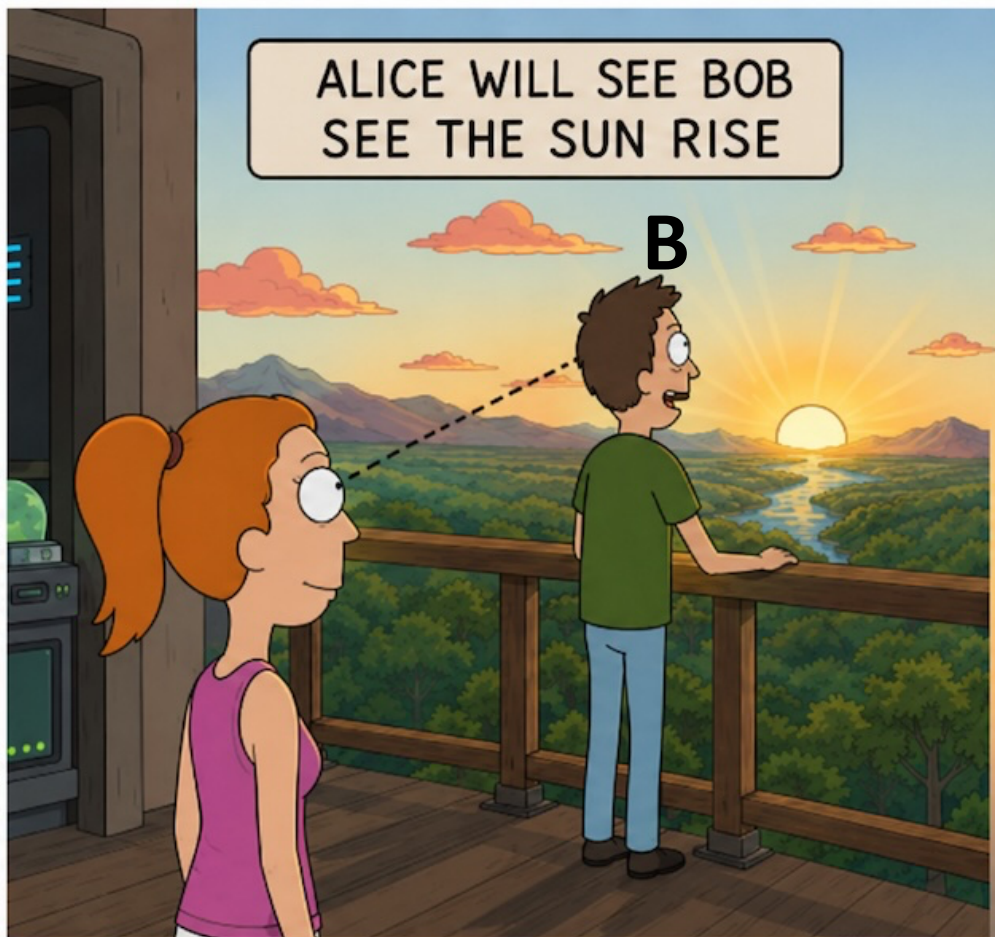
i.e. that $|\mathbf{P}(y|x_1, \dots, x_n) - \mu_W(y|x_1, \dots, x_n)| \longrightarrow 0$.

actual chances
according to univ. probability

chances as determined by
world W .

An emergent notion of **objective reality**

Suppose in **A**-world, there is another bit-string valued random variable, **B**.

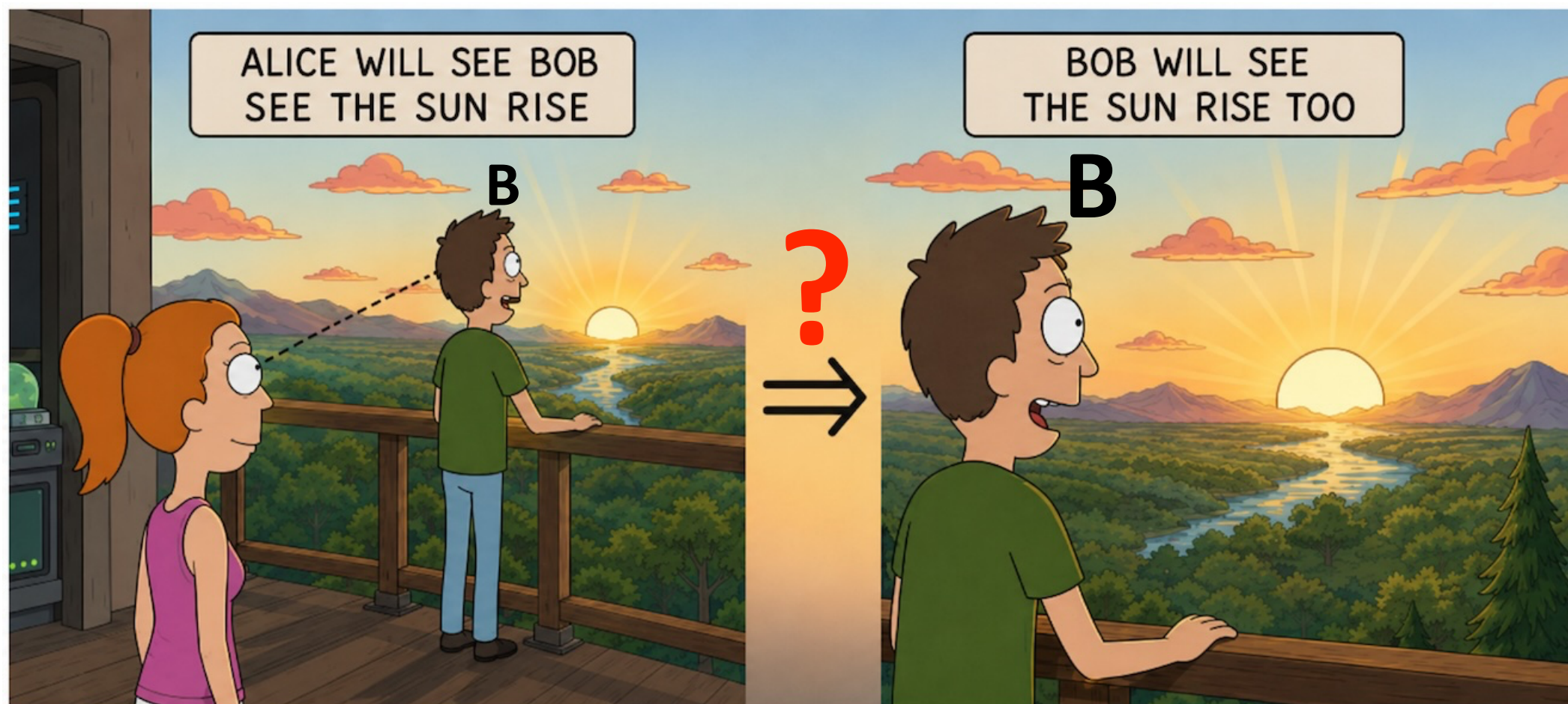


$$\mathbf{P}_{3\text{rd}}(y \mid x) \approx 1$$

An emergent notion of **objective reality**

Suppose in **A**-world, there is another bit-string valued random variable, **B**.

Does **B** faithfully represent some first-person perspective?



$$P_{3rd}(y | x) \approx 1$$

$$P_{1st}(y | x) \approx 1$$

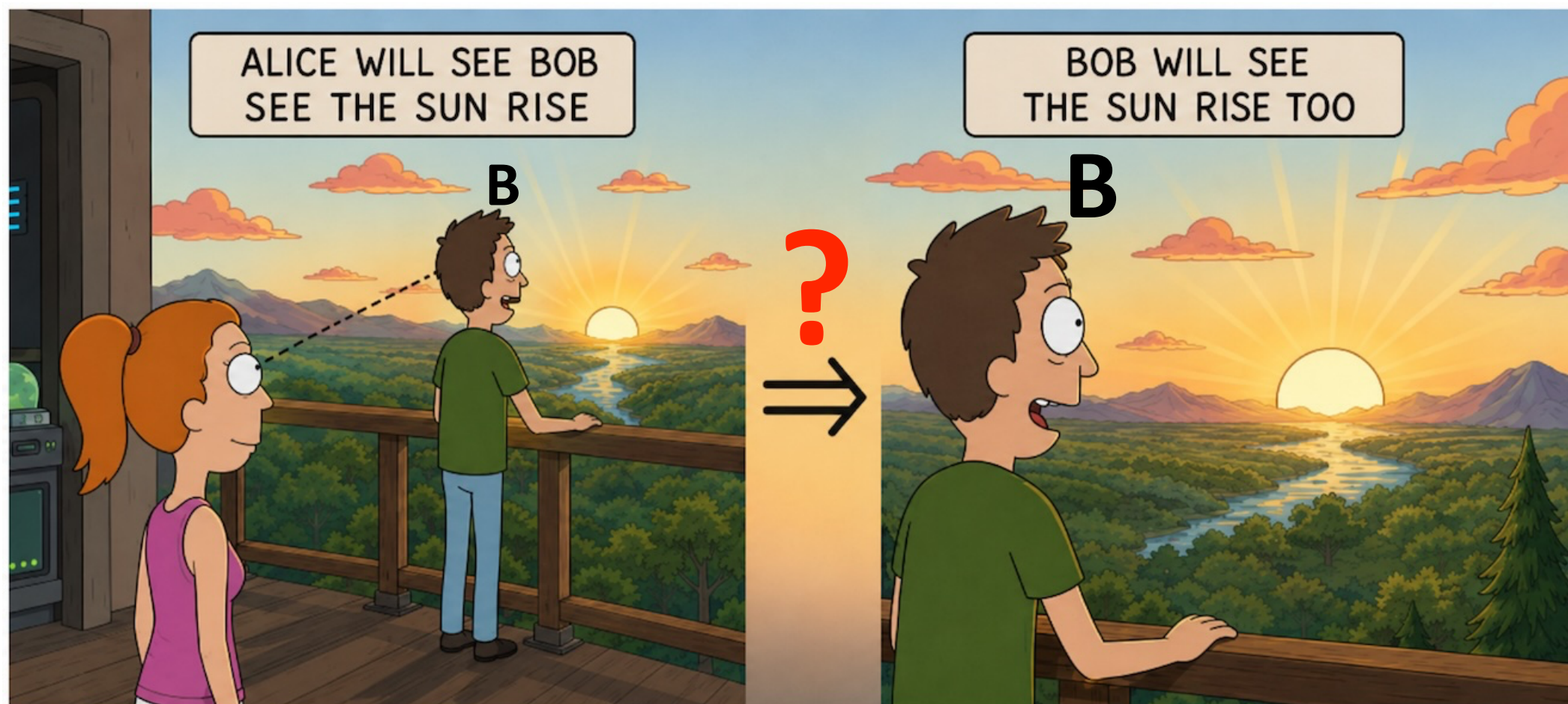
An emergent notion of **objective reality**

Suppose in **A**-world, there is another bit-string valued random variable, **B**.

Does **B** faithfully represent some first-person perspective?

P_{3rd} : how **B** changes over time according to the prob. laws of **A**-world.

P_{1st} : the actual “first-person” chances of **Bob’s** self state changes.



$$P_{3rd}(y | x) \approx 1$$

$$P_{1st}(y | x) \approx 1$$

An emergent notion of **objective reality**

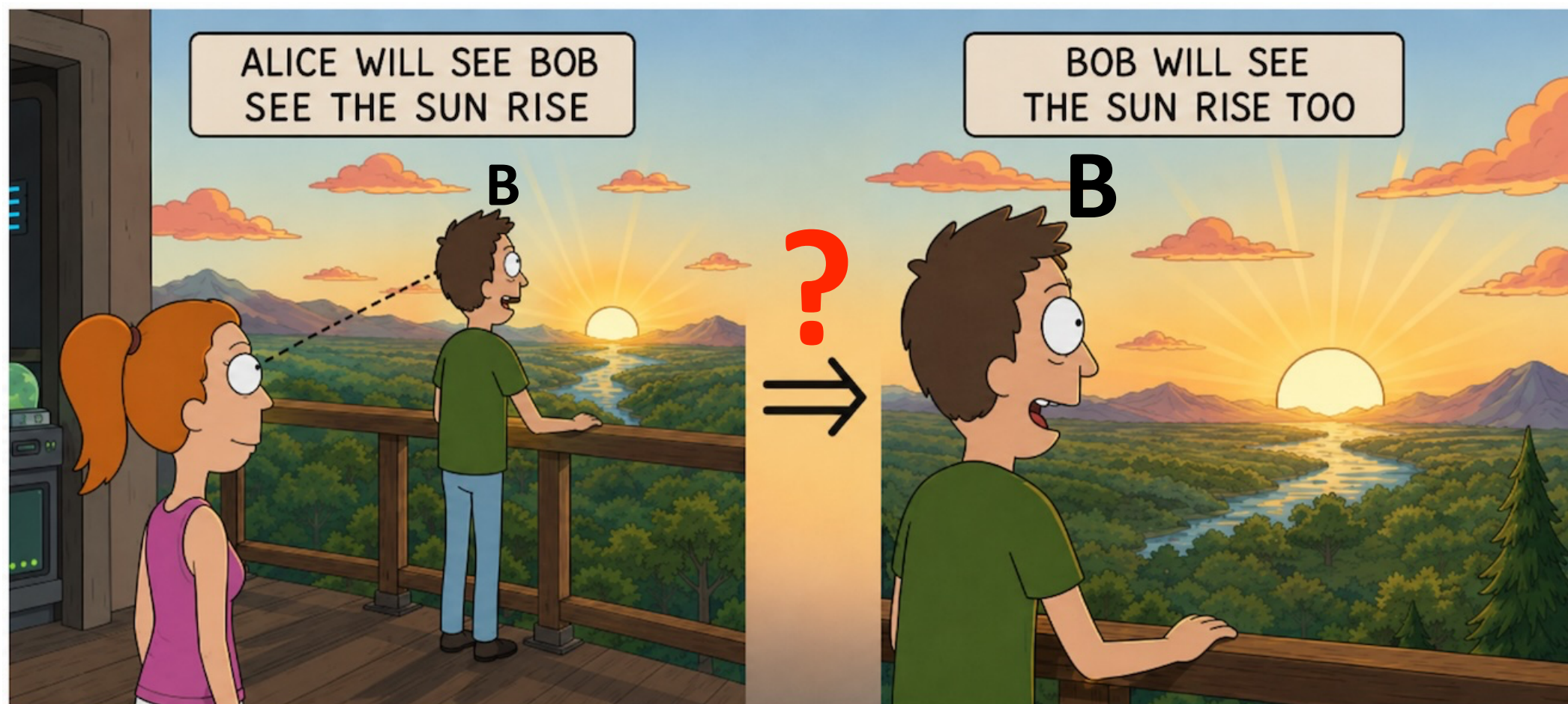
Suppose in **A**-world, there is another bit-string valued random variable, **B**.

Does **B** faithfully represent some first-person perspective?

P_{3rd} : how **B** changes over time according to the prob. laws of **A**-world.

P_{1st} : the actual “first-person” chances of **Bob’s** self state changes.

$P_{1st} = \text{universal probability } P.$



$$P_{3rd}(y | x) \approx 1$$

$$P_{1st}(y | x) \approx 1$$

An emergent notion of **objective reality**

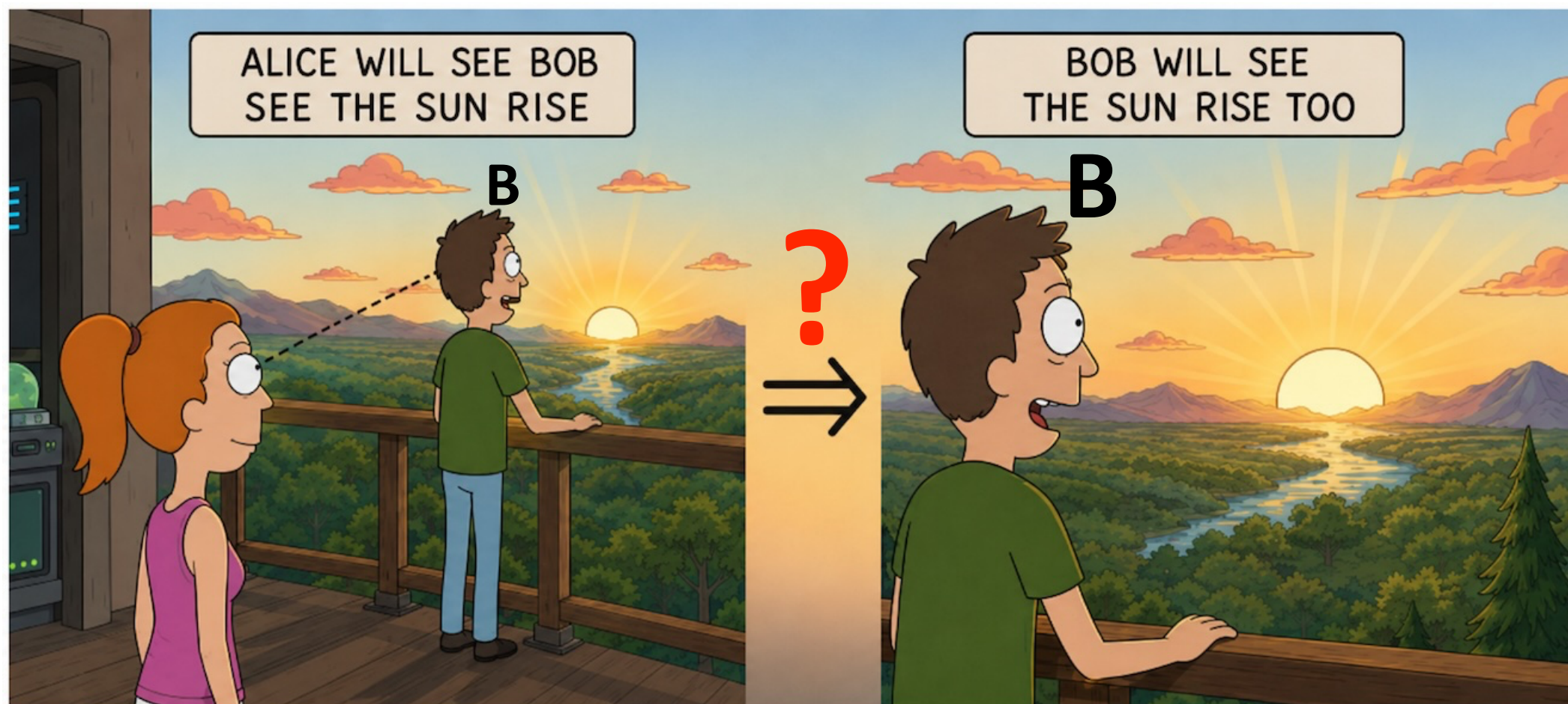
Suppose in **A**-world, there is another bit-string valued random variable, **B**.

Does **B** faithfully represent some first-person perspective?

P_{3rd} : how **B** changes over time according to the prob. laws of **A**-world.

P_{1st} : the actual “first-person” chances of **Bob**’s self state changes.

$P_{1st} = \text{universal probability } P.$ **Do they agree?**



$$P_{3rd}(y | x) \approx 1$$

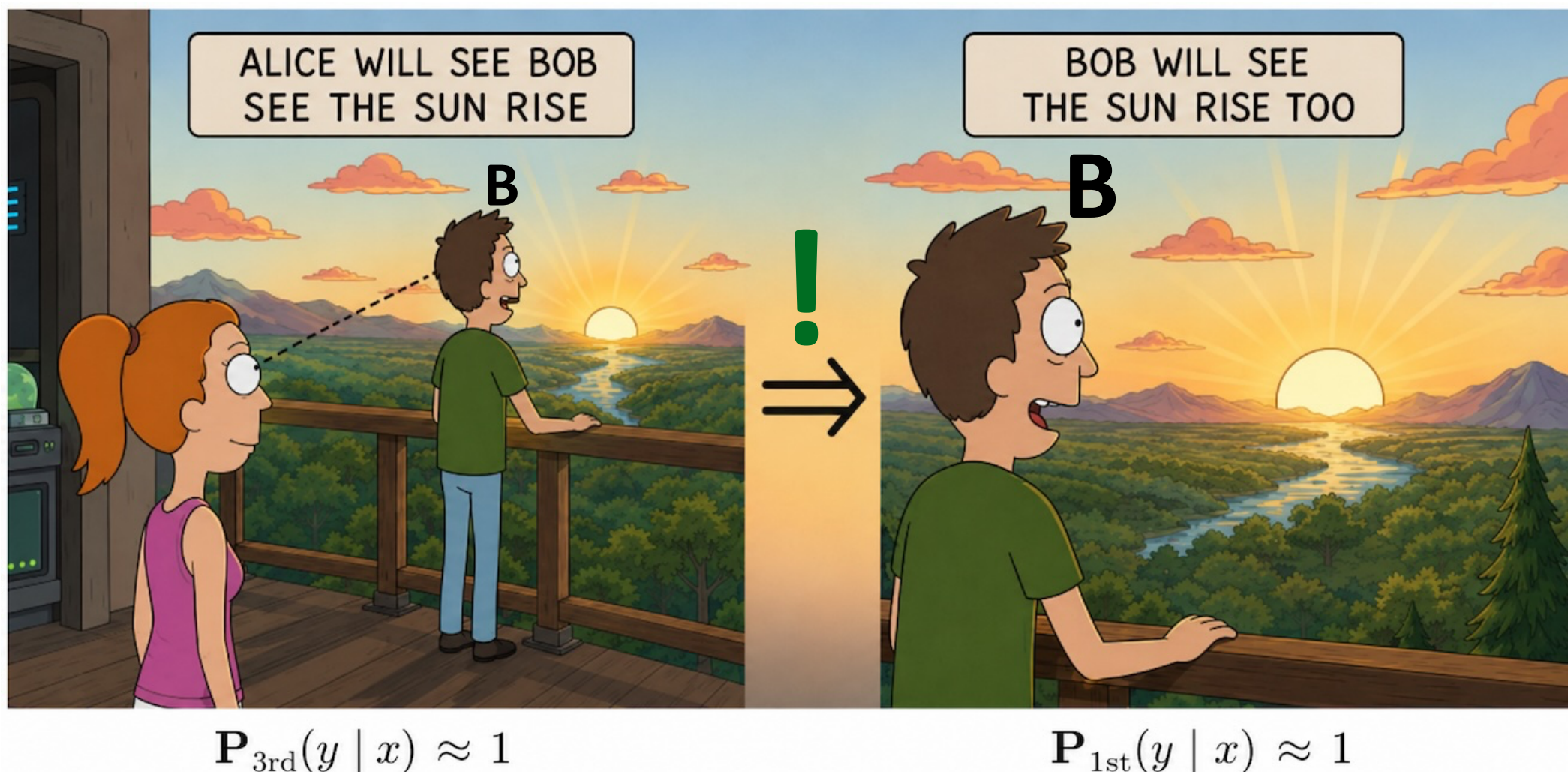
$$P_{1st}(y | x) \approx 1$$

An emergent notion of **objective reality**

Theorem. If **B** is a computable function of the state of **A**-world, then

$$|\mathbf{P}_{1\text{st}}(y|x_1, \dots, x_n) - \mathbf{P}_{3\text{rd}}(y|x_1, \dots, x_n)| \xrightarrow{n \rightarrow \infty} 0,$$

so asymptotically, **we have the consistency that we expect.**



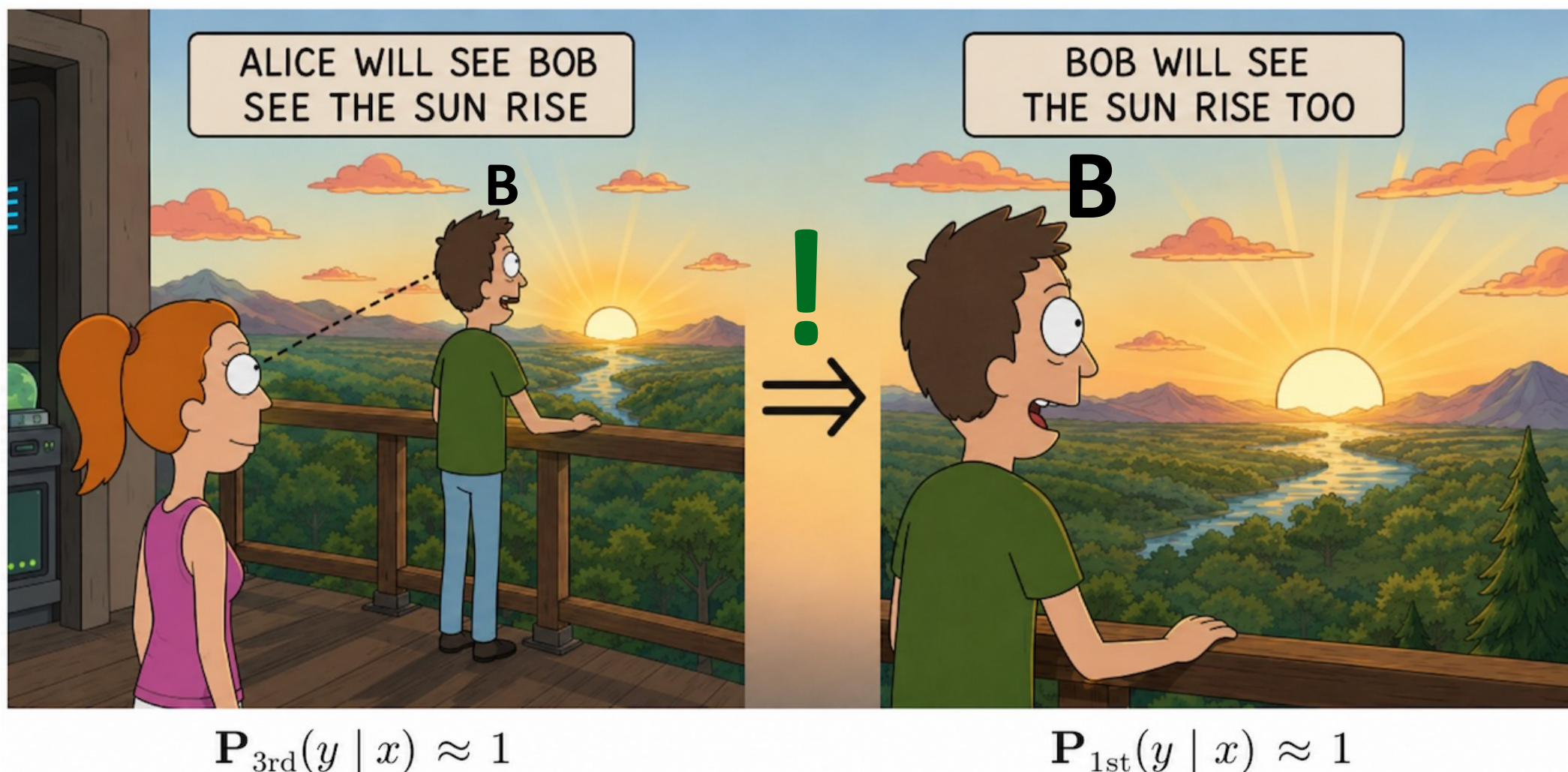
An emergent notion of **objective reality**

Theorem. If **B** is a computable function of the state of **A**-world, then

$$|\mathbf{P}_{1\text{st}}(y|x_1, \dots, x_n) - \mathbf{P}_{3\text{rd}}(y|x_1, \dots, x_n)| \xrightarrow{n \rightarrow \infty} 0,$$

so asymptotically, **we have the consistency that we expect.**

In fact, $\mathbf{P}_{1\text{st}}(y|x) \approx \mathbf{P}_{3\text{rd}}(y|x)$ if $K(x) \gg K(\mathbf{P}_{3\text{rd}})$, i.e. if Bob's self state is complex compared to its statistical **A**-world laws.

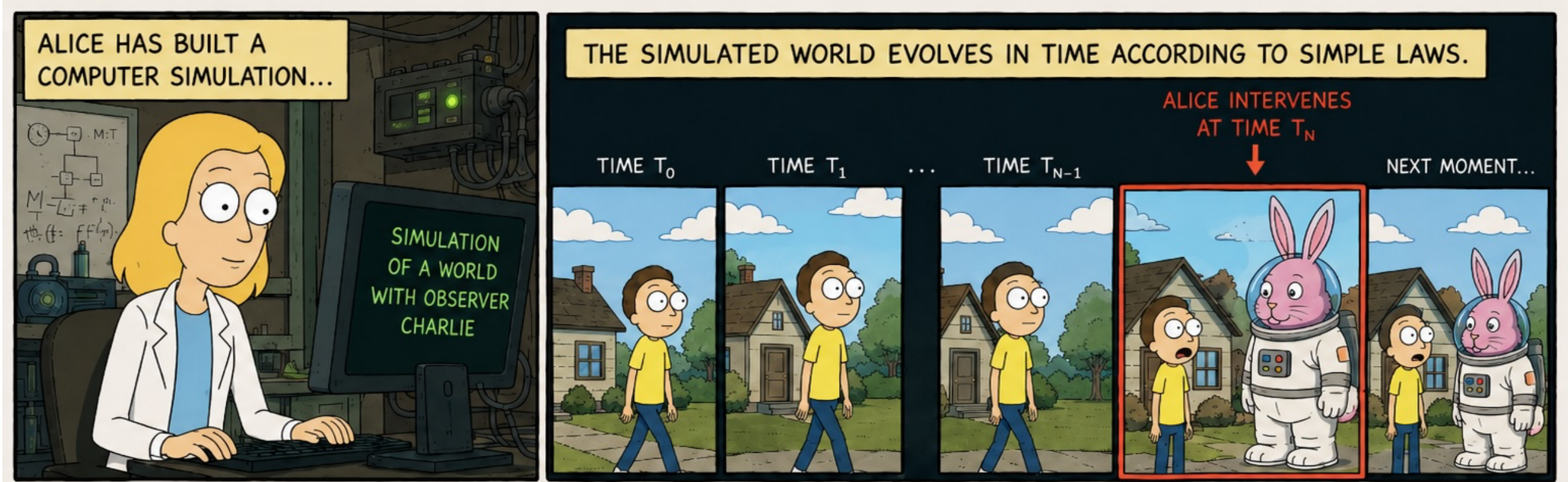


Probabilistic changelings

But if $K(x) \ll K(\mathbf{P}_{3\text{rd}})$, we can have situations where $\mathbf{P}_{3\text{rd}} \not\approx \mathbf{P}_{1\text{st}}$.

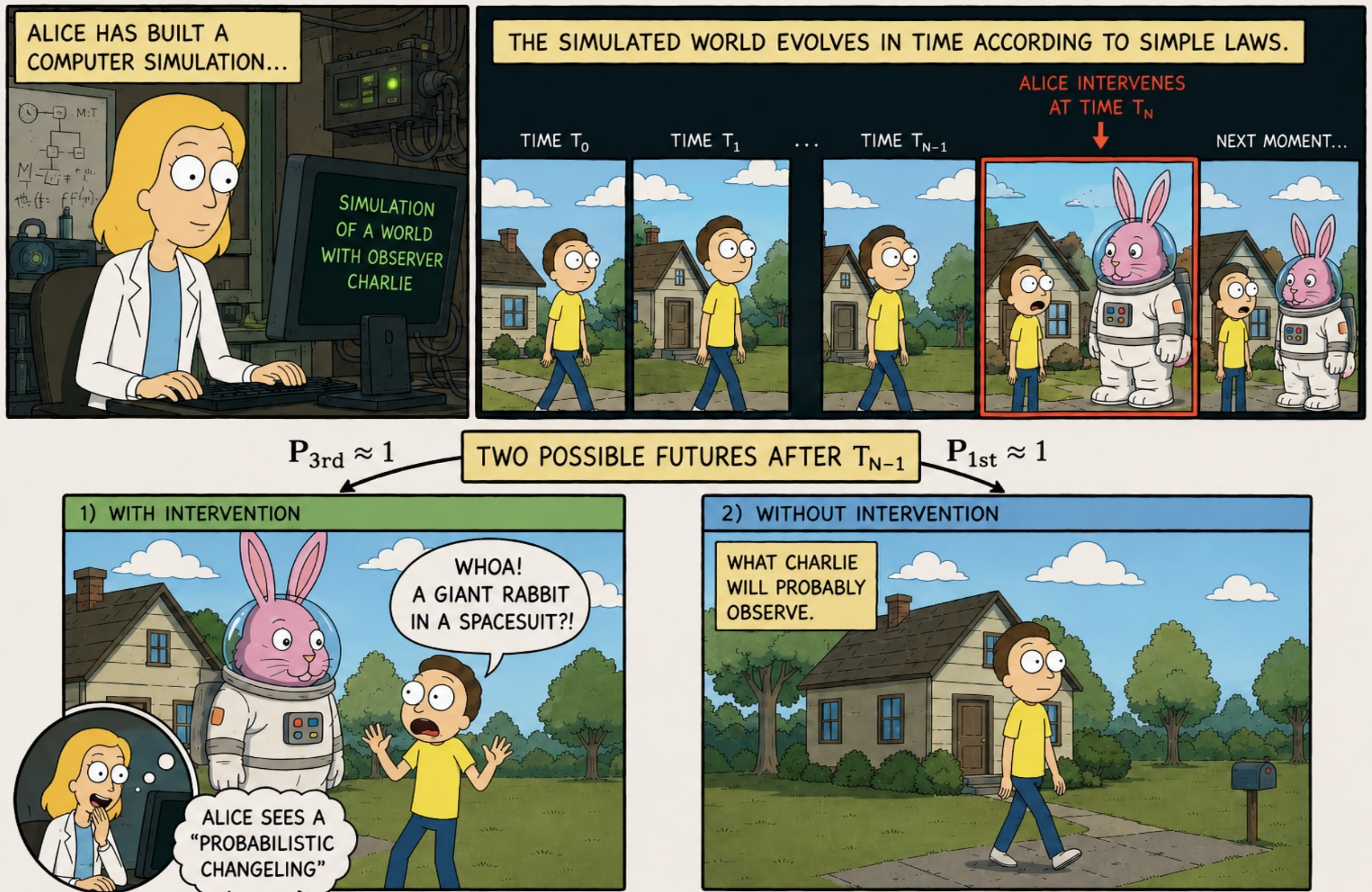
Probabilistic changelings

But if $K(x) \ll K(\mathbf{P}_{3\text{rd}})$, we can have situations where $\mathbf{P}_{3\text{rd}} \not\approx \mathbf{P}_{1\text{st}}$.



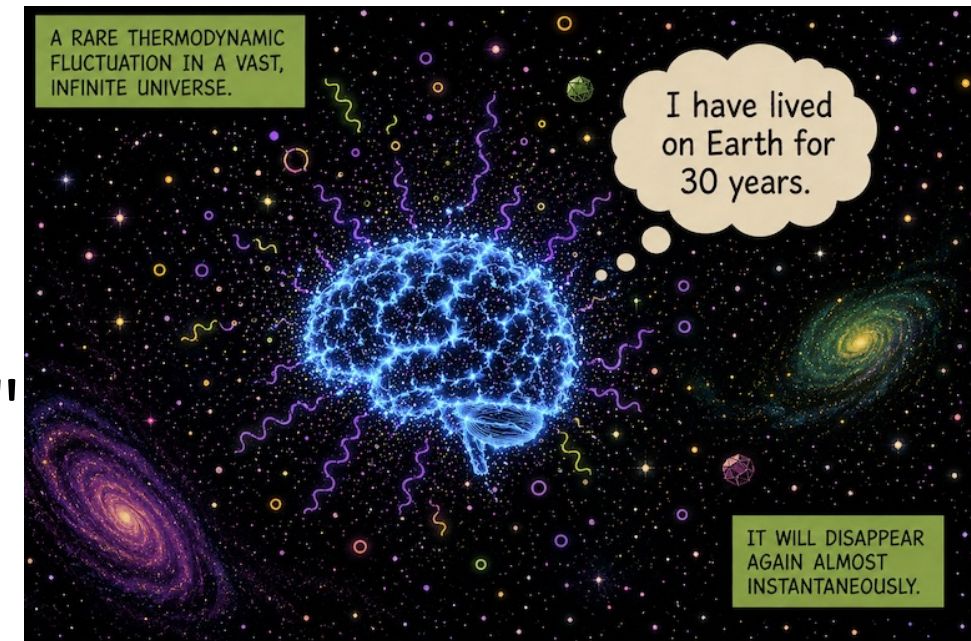
Probabilistic changelings

But if $K(x) \ll K(\mathbf{P}_{3\text{rd}})$, we can have situations where $\mathbf{P}_{3\text{rd}} \not\approx \mathbf{P}_{1\text{st}}$.



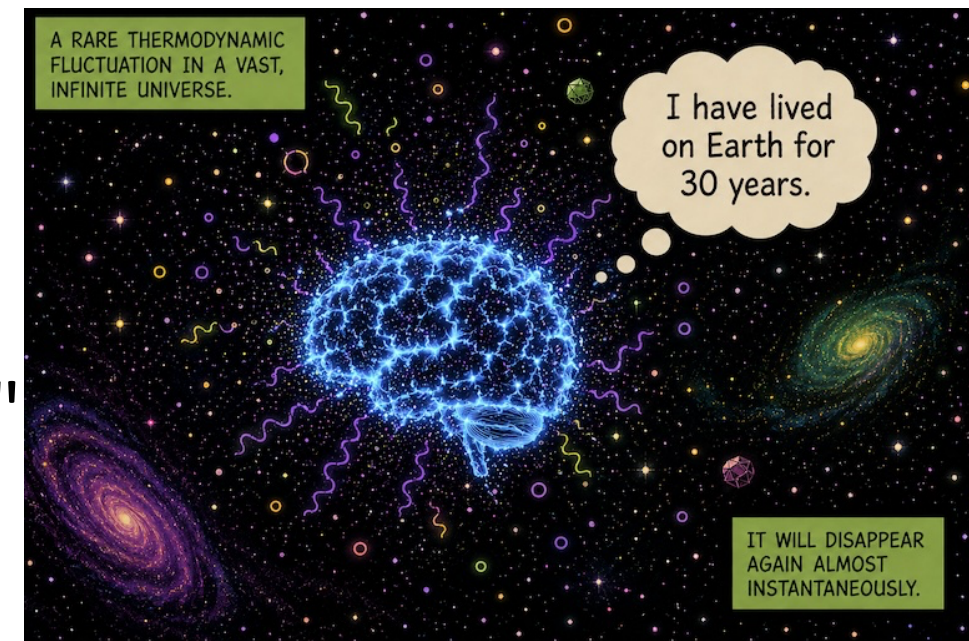
Predictions of Algorithmic Idealism

- Even if there are many **Boltzmann brains** in your emergent external world, "business as usual" is always more probable. In particular, cosmologists cannot "count BBs" to rule out some cosmological models.



Predictions of Algorithmic Idealism

- Even if there are many **Boltzmann brains** in your emergent external world, "business as usual" is always more probable. In particular, cosmologists cannot "count BBs" to rule out some cosmological models.



- **Most of our intuitive metaphysical assumptions are demolished.** For example, all self states "exist" (and the conscious ones are conscious) *regardless* of whether they are actualized or realized in some "world". The notion of "world" and mechanistic explanations is a mere statistical approximation and emergent phenomenon. There is no fundamental distinction between "base reality" and "simulation"; reality is fundamentally immaterial / mathematical. Saying all this is an **unintended side effect of A.Id.**, which was merely built to **predict** all private experiments, *deliberately setting aside ontology*.

Predictions of Algorithmic Idealism - **beyond the bit model**

- Recall simulated Charlie who "leaks out of the simulation"?
In some models of A.Id., **death** is described in a very similar way. The Parfit graph model predicts that everyone does an infinite random walk on all self states, "restarting" life in different times and mathematical universes (in some of them on average more often than in others).



relative
probabilistic
changelings?

Predictions of Algorithmic Idealism - **beyond the bit model**

- Recall simulated Charlie who "leaks out of the simulation"?
In some models of A.Id., **death** is described in a very similar way. The Parfit graph model predicts that everyone does an infinite random walk on all self states, "restarting" life in different times and mathematical universes (in some of them on average more often than in others).



relative
probabilistic
changelings?

- In A.Id., the external world fundamentally supervenes on the self state. Hence, resetting the self state resets the external world to its earlier state, including all clocks and external records.
This leads to unexpected statistical phenomena, including the **violation of Bell inequalities despite signal locality**. Work in progress.

Summary

Summary

- Physics, Philosophy and future technology (presently with AI) have brought us enigmatic **private** (thought) **experiments** that cannot be resolved by means of the "standard" third-person view (**Restriction A**).
- **Algorithmic Idealism** is a response to this. It sets ontological questions aside, and its sole goal is to obtain predictions for private experiments. It is a framework that admits several possible mathematical **models**.
- In contrast to earlier versions of idealism, the (apparent) existence of a lawlike **external world** is not a miracle but a **prediction**. The elements of **objective reality** that we take for granted are not assumed, but proven as theorems under some assumptions.
- Despite its agreement with physics in standard situations, it implies a complete shift of perspective on topics such as the nature of reality, the Boltzmann brain problem, the simulation hypothesis, or death.

Summary

- Physics, Philosophy and future technology (presently with AI) have brought us enigmatic **private** (thought) **experiments** that cannot be resolved by means of the "standard" third-person view (**Restriction A**).
- **Algorithmic Idealism** is a response to this. It sets ontological questions aside, and its sole goal is to obtain predictions for private experiments. It is a framework that admits several possible mathematical **models**.
- In contrast to earlier versions of idealism, the (apparent) existence of a lawlike **external world** is not a miracle but a **prediction**. The elements of **objective reality** that we take for granted are not assumed, but proven as theorems under some assumptions.
- Despite its agreement with physics in standard situations, it implies a complete shift of perspective on topics such as the nature of reality, the Boltzmann brain problem, the simulation hypothesis, or death.

Thank you! mpmueller.net/aid