

Thinking twice inside the box: is Wigner's Friend really about Quantum Theory?

Markus P. Müller

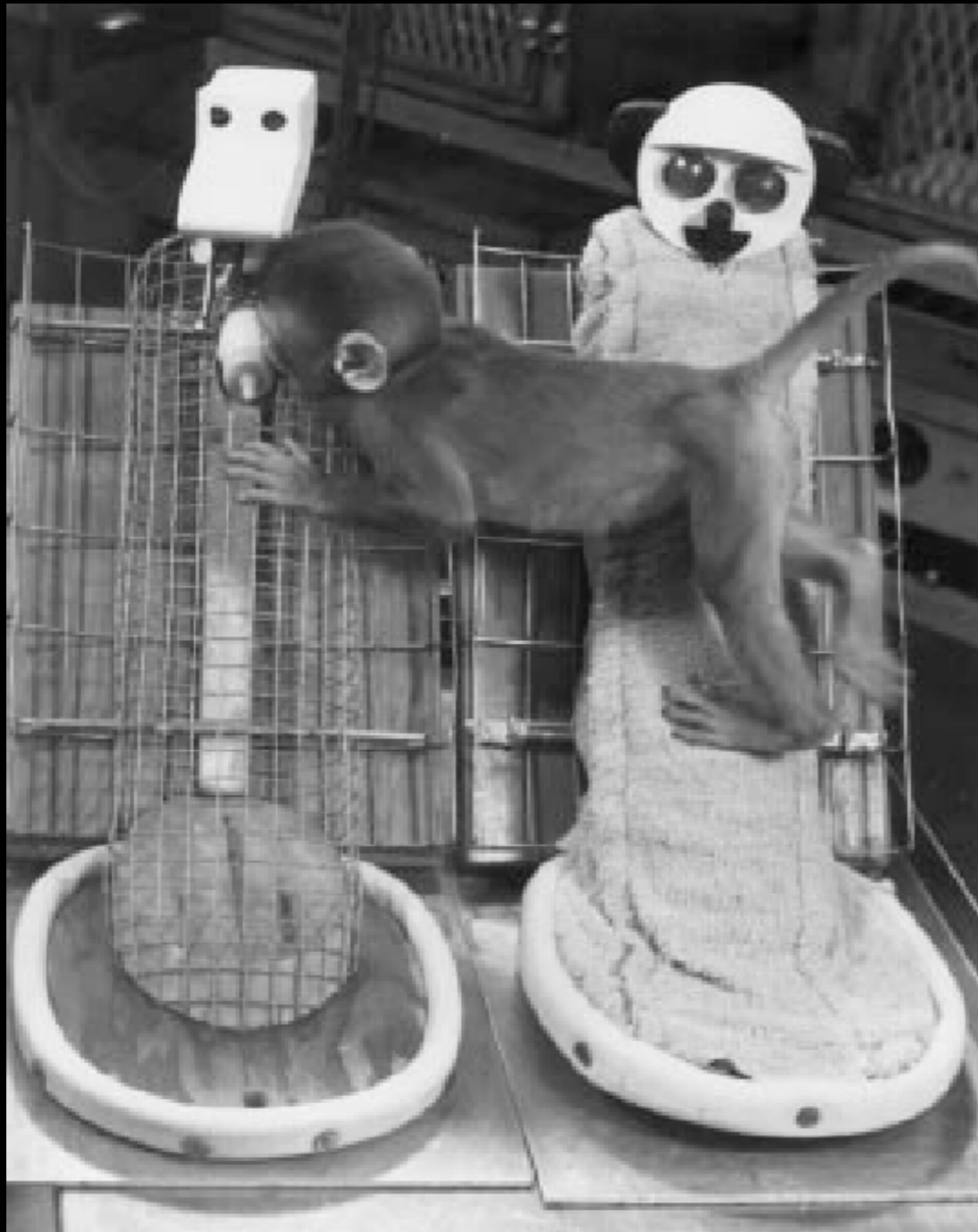
Institute for Quantum Optics and Quantum Information (IQOQI), Vienna
Perimeter Institute for Theoretical Physics (PI), Waterloo, Canada



Behaviorism lacks predictive power

Behaviorism lacks predictive power

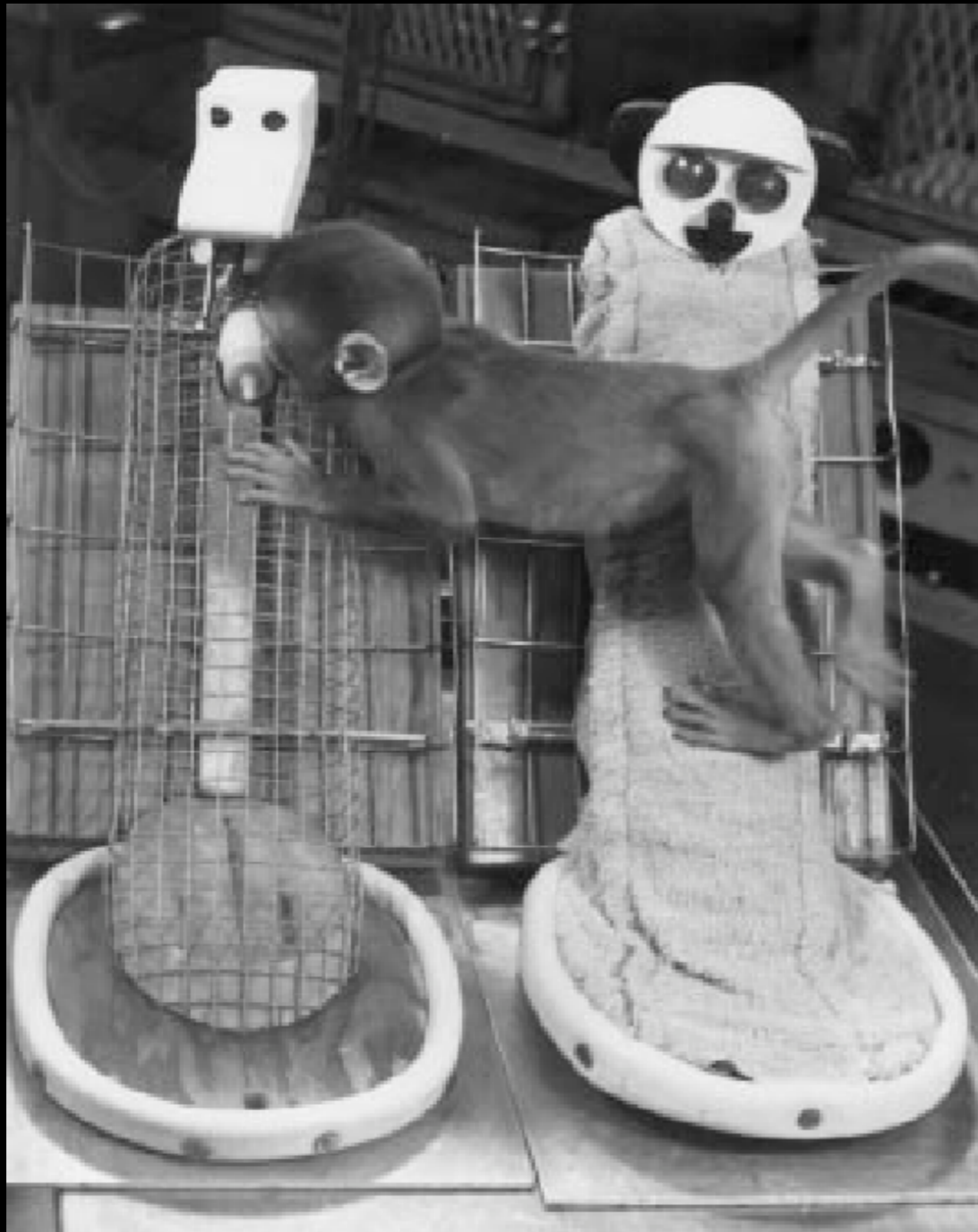
Harry Harlow, 1957: rhesus infant experiments



“The monkeys overwhelmingly chose the cloth mother, with or without food, only visiting the wire mother that had food when needing sustenance.”

Behaviorism lacks predictive power

Harry Harlow, 1957: rhesus infant experiments



“The monkeys overwhelmingly chose the cloth mother, with or without food, only visiting the wire mother that had food when needing sustenance.”

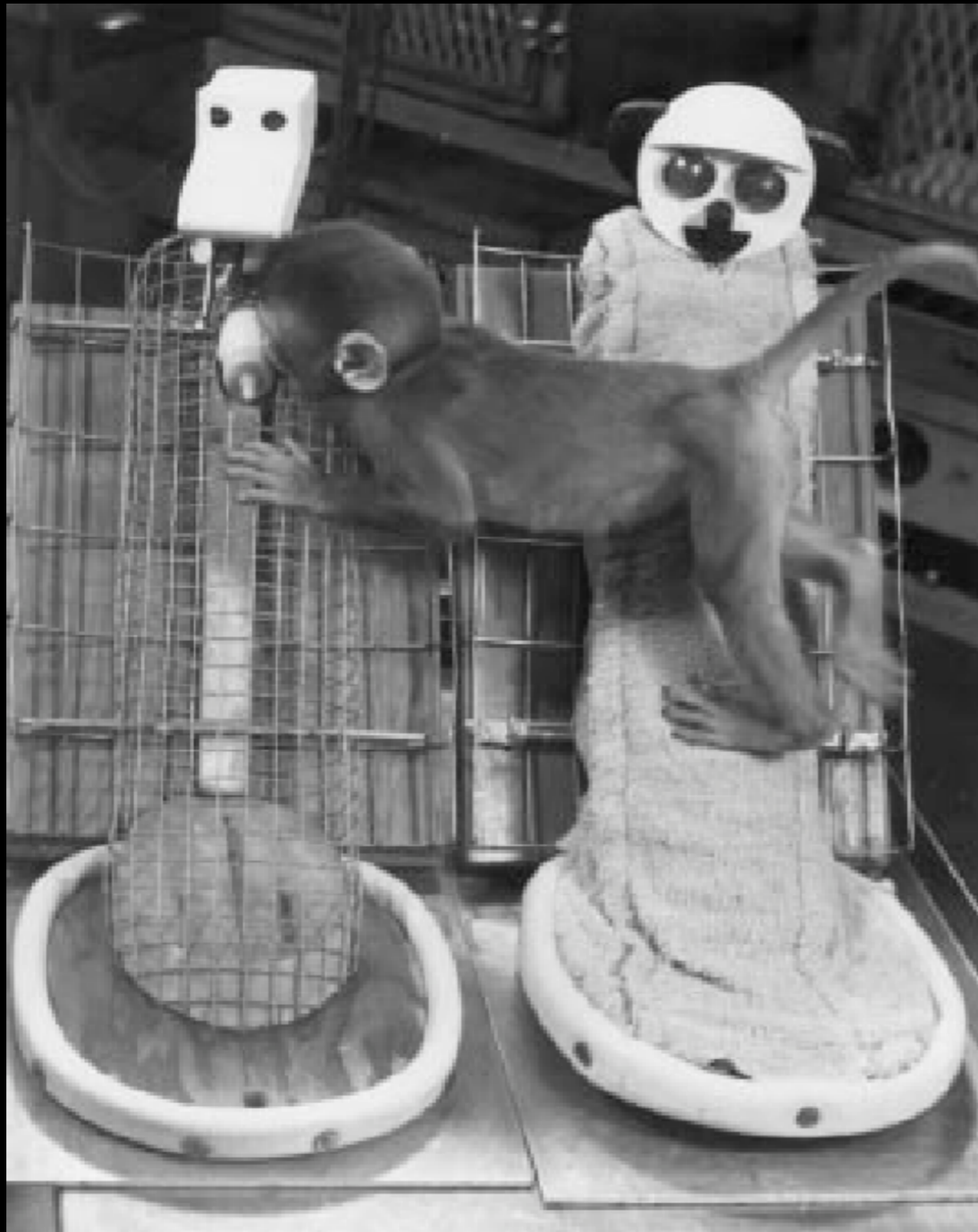
“**Behaviorism** is a systematic approach to understand the behavior of humans and other animals.^{[1][2]} It assumes that behavior is either a reflex elicited by the pairing of certain antecedent stimuli in the environment, or a consequence of that individual's history, especially including reinforcement and punishment contingencies, [...]”

“The cognitive revolution of the late 20th century largely replaced behaviorism as an explanatory theory with cognitive psychology, which unlike behaviorism views internal mental states as explanations for observable behavior.”

Harlow “described his experiments as a study of love.”

Behaviorism lacks predictive power

Harry Harlow, 1957: rhesus infant experiments



“The monkeys overwhelmingly chose the cloth mother, with or without food, only visiting the wire mother that had food when needing sustenance.”

“**Behaviorism** is a systematic approach to understand the behavior of humans and other animals.^{[1][2]} It assumes that behavior is either a reflex elicited by the pairing of certain antecedent stimuli in the environment, or a consequence of that individual's history, especially including reinforcement and punishment contingencies, [...]”

“The cognitive revolution of the late 20th century largely replaced behaviorism as an explanatory theory with cognitive psychology, which unlike behaviorism views internal mental states as explanations for observable behavior.”

Harlow “described his experiments as a study of love.”

Physics is still “behaviorist” — the external view lacks predictive power.

Is Wigner's Friend really about Quantum Theory?

Is Wigner's Friend really about Quantum Theory?

Claim: **No**, it has a much broader significance.

It is a particular example of

Restriction A: Our physical theories do not (sometimes *cannot*) give us joint predictions for the future observations of all **Agents**.

Is Wigner's Friend really about Quantum Theory?

Claim: **No**, it has a much broader significance.

It is a particular example of

Restriction A: Our physical theories do not (sometimes *cannot*) give us joint predictions for the future observations of all **Agents**.

Sometimes even for *single* agents.

This is relative to some theory T and background assumptions.

Is Wigner's Friend really about Quantum Theory?

Claim: **No**, it has a much broader significance.

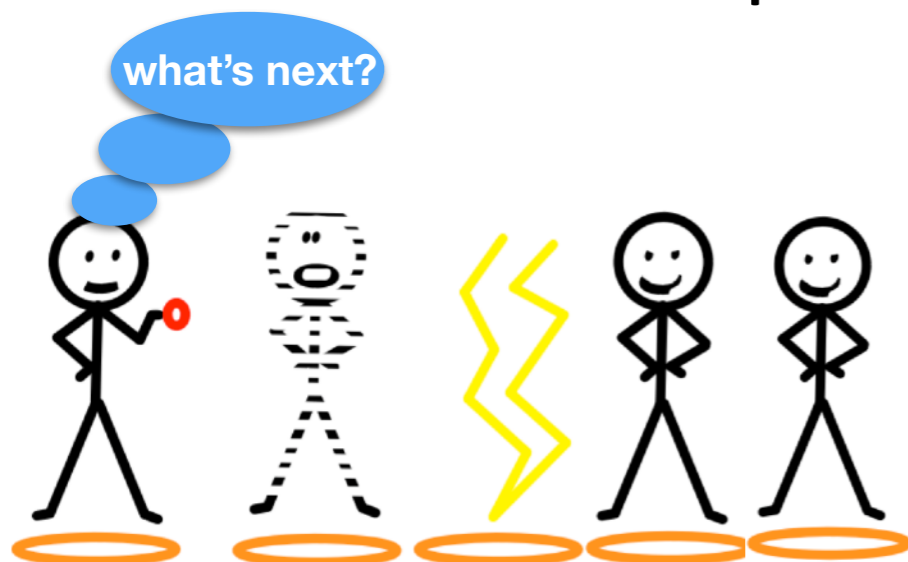
It is a particular example of

Restriction A: Our physical theories do not (sometimes *cannot*) give us joint predictions for the future observations of all **Agents**.

Sometimes even for *single* agents.

This is relative to some theory T and background assumptions.

Further instances **beyond quantum physics:** Duplication scenarios, Boltzmann brain problem, computer simulation of agents, death.



Is Wigner's Friend really about Quantum Theory?

Claim: **No**, it has a much broader significance.

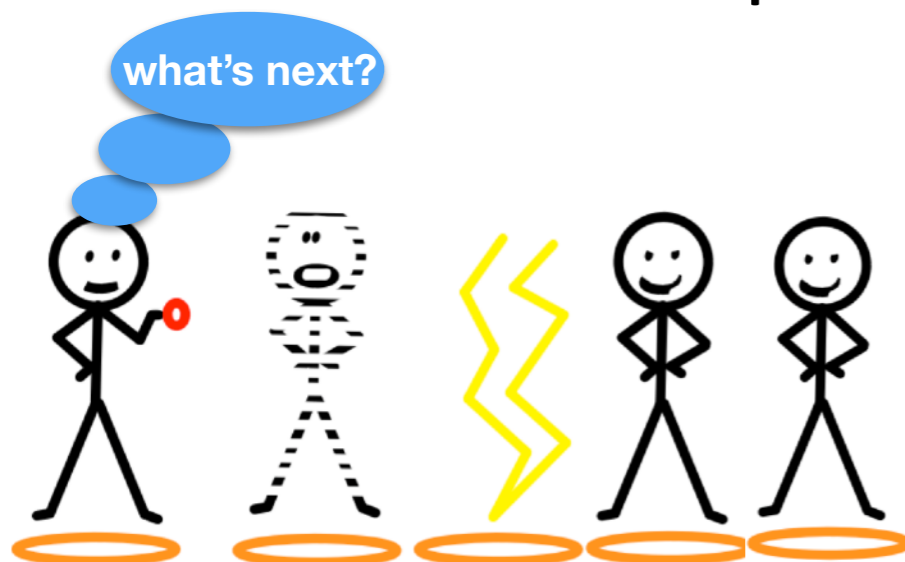
It is a particular example of

Restriction A: Our physical theories do not (sometimes *cannot*) give us joint predictions for the future observations of all **Agents**.

Sometimes even for *single* agents.

This is relative to some theory T and background assumptions.

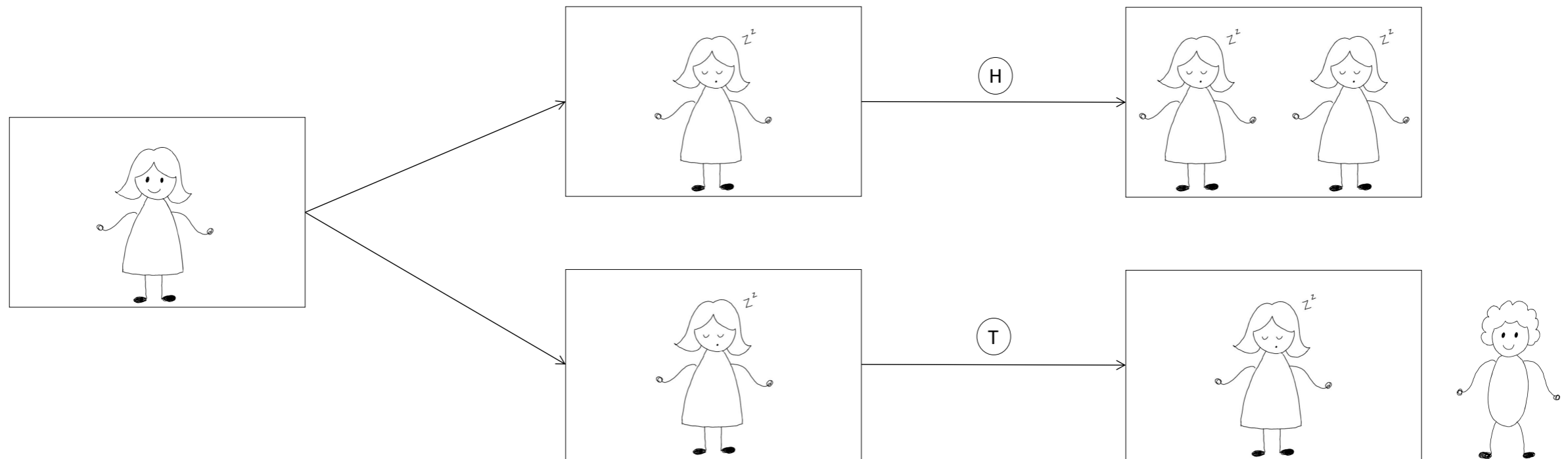
Further instances **beyond quantum physics**: Duplication scenarios, Boltzmann brain problem, computer simulation of agents, death.



Consequence: need **idealist/fragmentalist** approach, predicting for each *single* agent what they should believe to experience next.

Overview

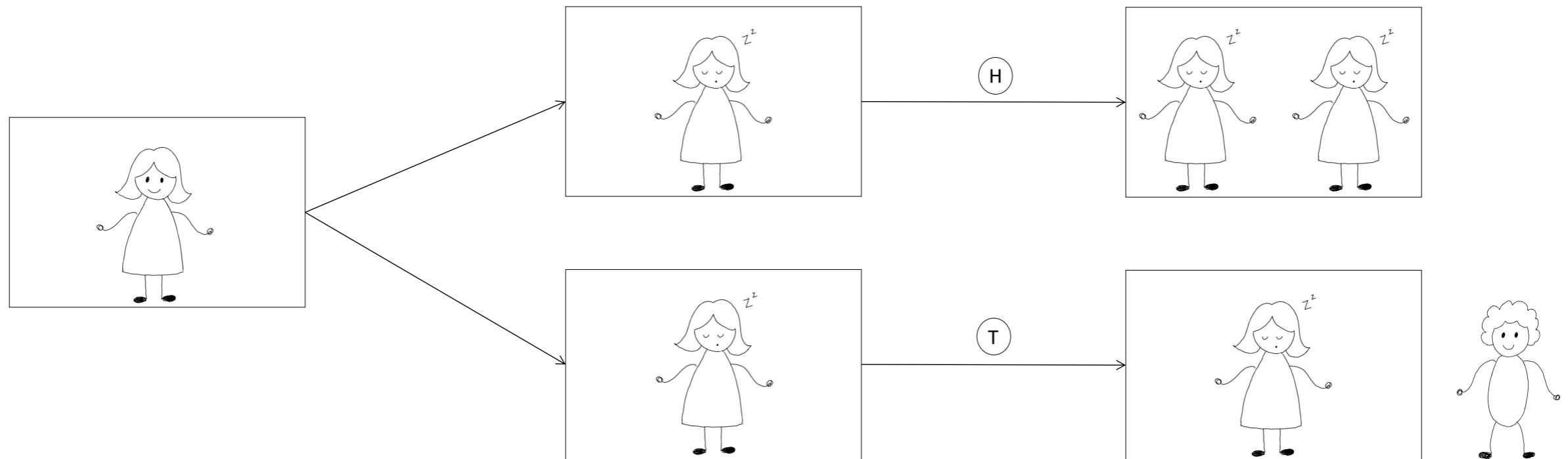
1. Reproducing WF phenomenology with classical **duplication** (“thinking twice inside the box”)



2. A common *structural* core: **Restriction A**
3. Restriction A elsewhere: **Boltzmann brain problem**
4. Consequence: **Fragmentalism/idealism**

Overview

1. Reproducing WF phenomenology with classical **duplication** (“thinking twice inside the box”)



2. A common *structural* core: **Restriction A**

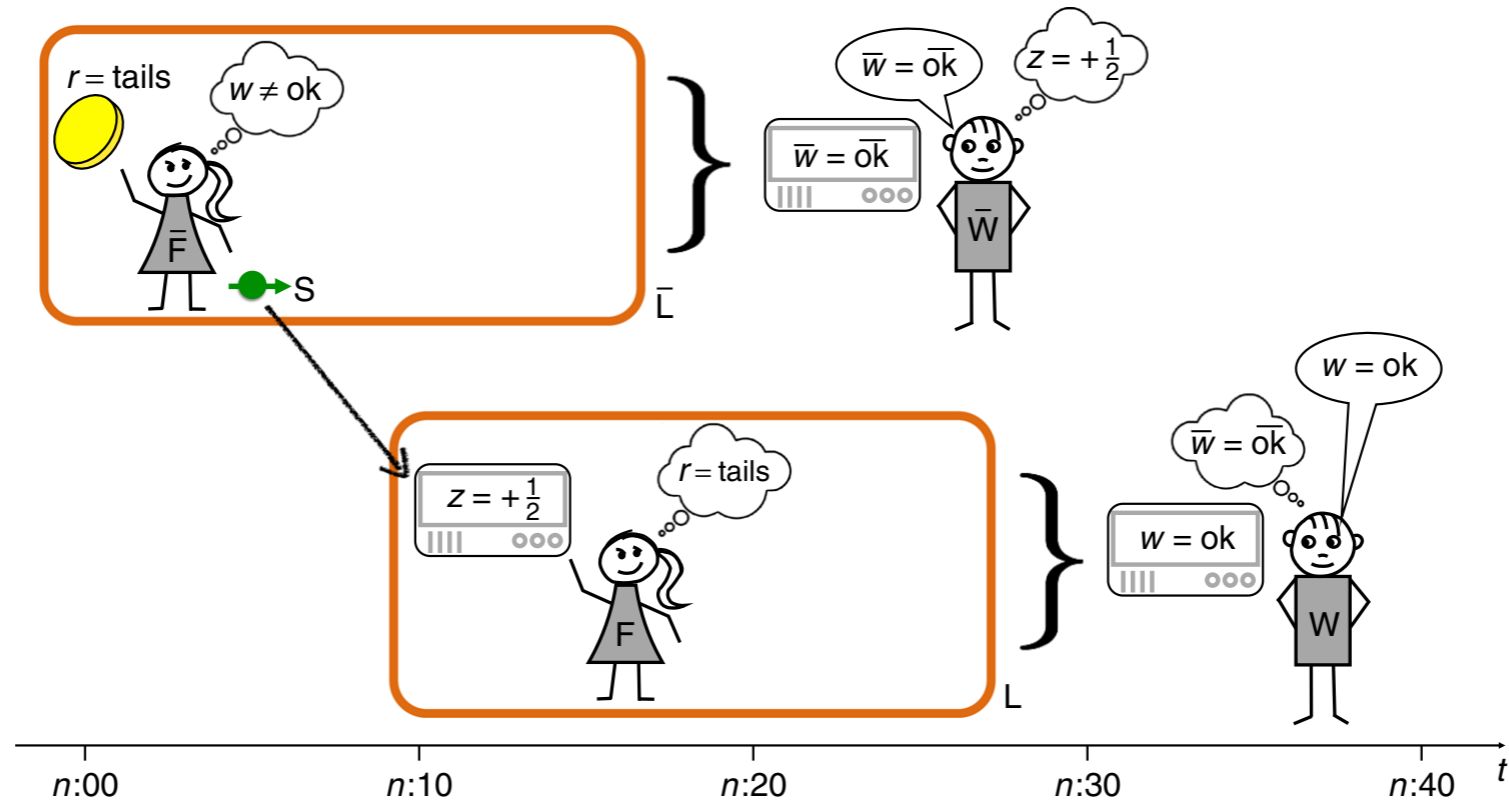
3. Restriction A elsewhere: **Boltzmann brain problem**

4. Consequence: **Fragmentalism/idealism**

The Frauchiger-Renner Scenario

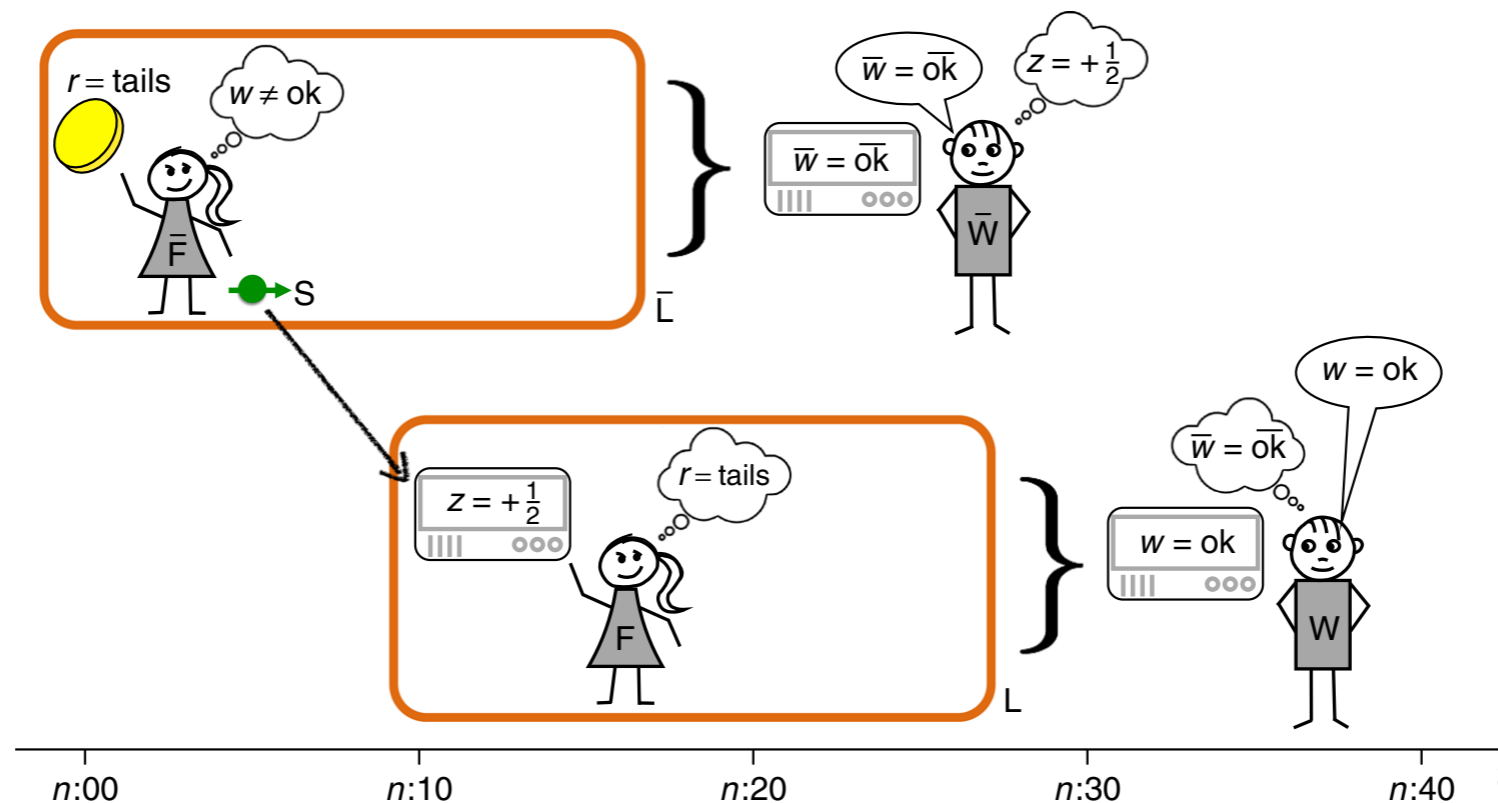
The Frauchiger-Renner Scenario

D. Frauchiger and R. Renner, Nat. Commun. **9**, 3711 (2018).



The Frauchiger-Renner Scenario

D. Frauchiger and R. Renner, Nat. Commun. **9**, 3711 (2018).

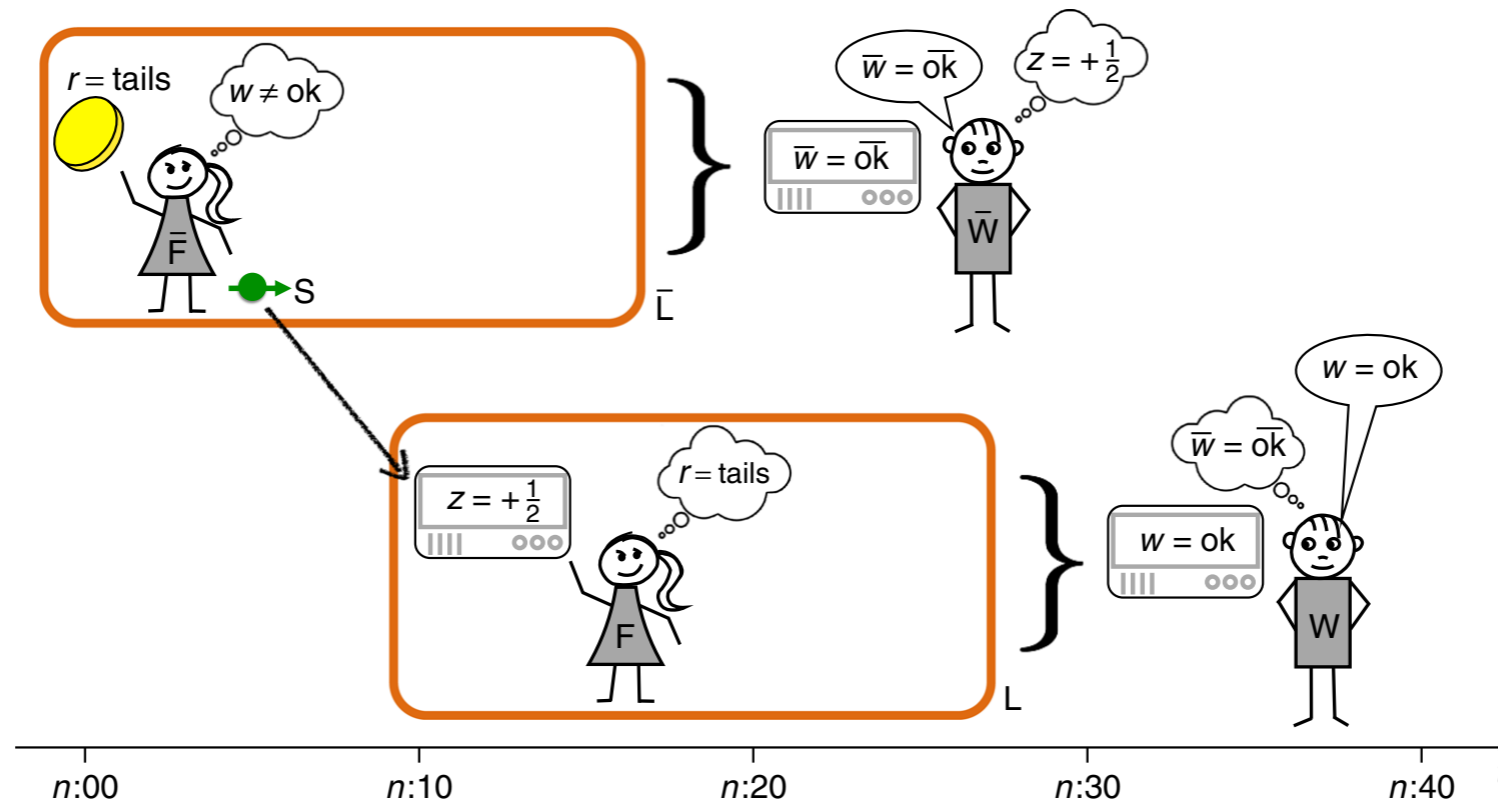


Incompatibility of three assumptions:

- (Q): Quantum theory is universally valid.
- (S): Measurement outcomes must be single-valued.
- (C): The predictions of different agents are consistent.

The Frauchiger-Renner Scenario

D. Frauchiger and R. Renner, Nat. Commun. **9**, 3711 (2018).



Incompatibility of three assumptions:

- (Q): Quantum theory is universally valid.
- (S): Measurement outcomes must be single-valued.
- **(C): The predictions of different agents are consistent.**

The Frauchiger-Renner Scenario

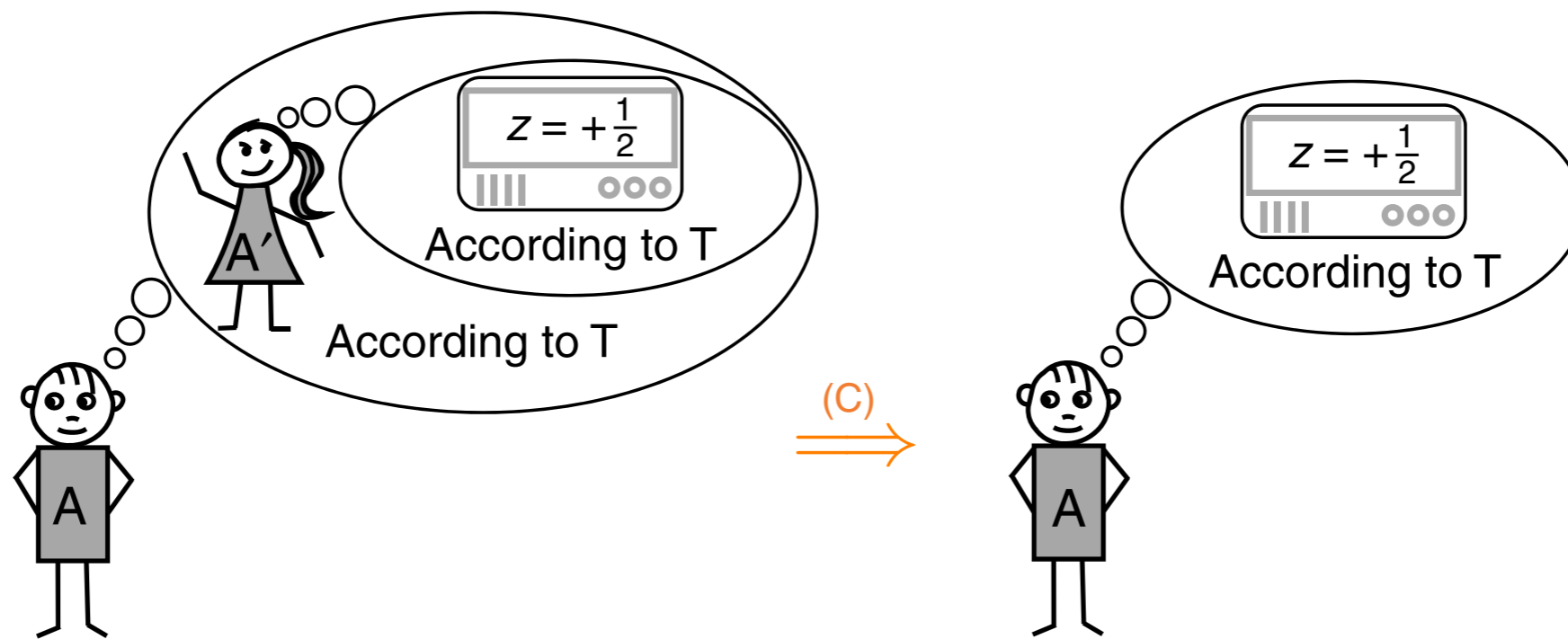
Incompatibility of three assumptions:

- (Q): Quantum theory is universally valid.
- (S): Measurement outcomes must be single-valued.
- **(C): The predictions of different agents are consistent.**

The Frauchiger-Renner Scenario

Incompatibility of three assumptions:

- (Q): Quantum theory is universally valid.
- (S): Measurement outcomes must be single-valued.
- **(C): The predictions of different agents are consistent.**



Box 3: Assumption (C)

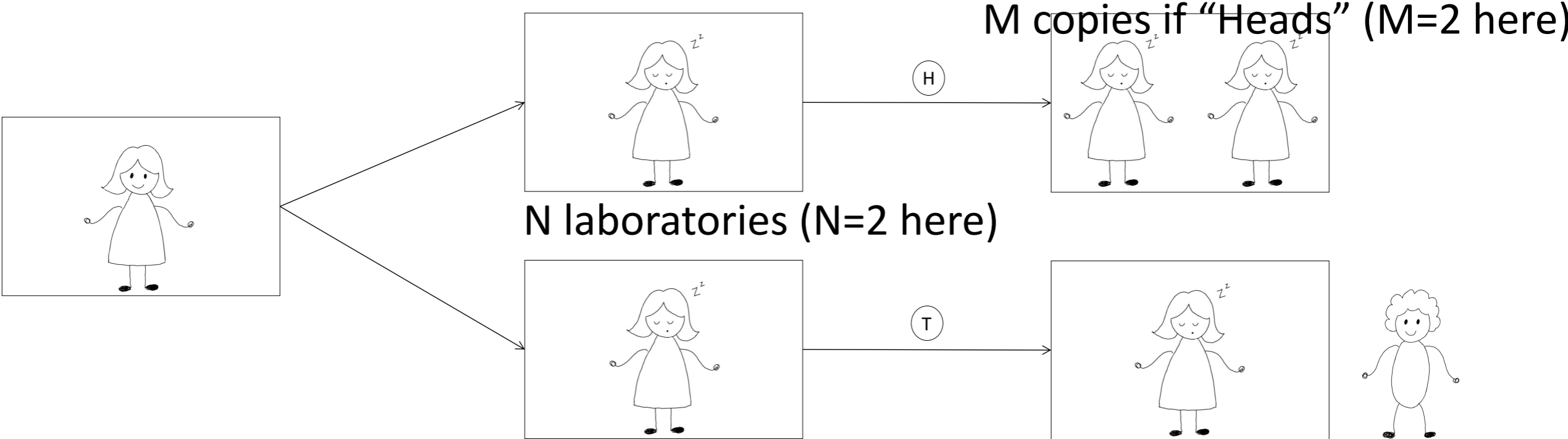
Suppose that agent A has established that

Statement A⁽ⁱ⁾: "I am certain that agent A', upon reasoning within the same theory as the one I am using, is certain that $x = \xi$ at time t ."

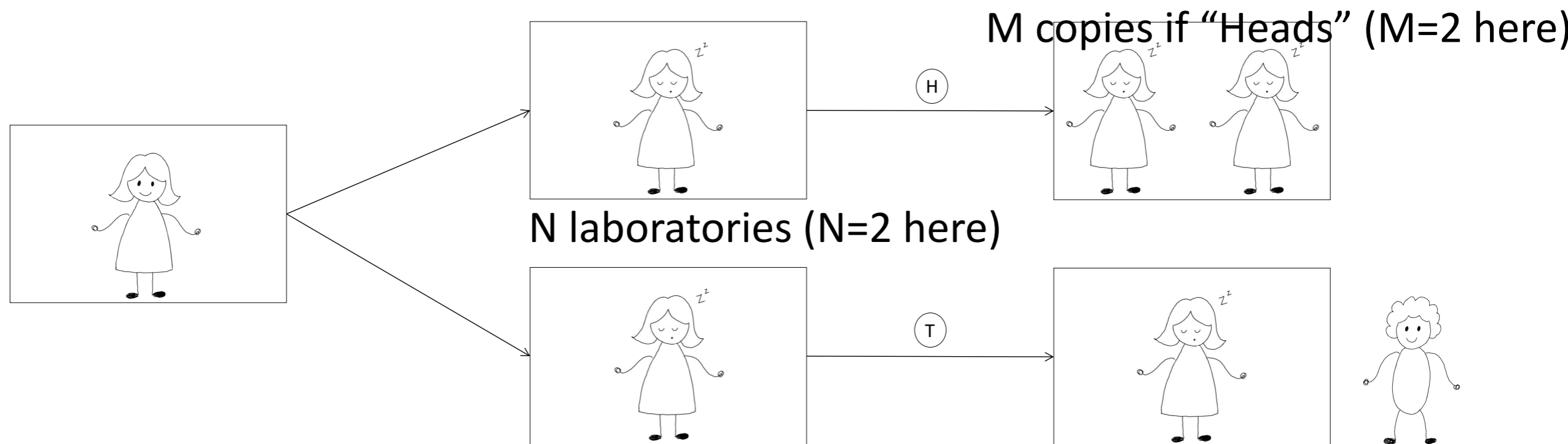
Then agent A can conclude that

Statement A⁽ⁱⁱ⁾: "I am certain that $x = \xi$ at time t ."

A classical thought experiment



A classical thought experiment

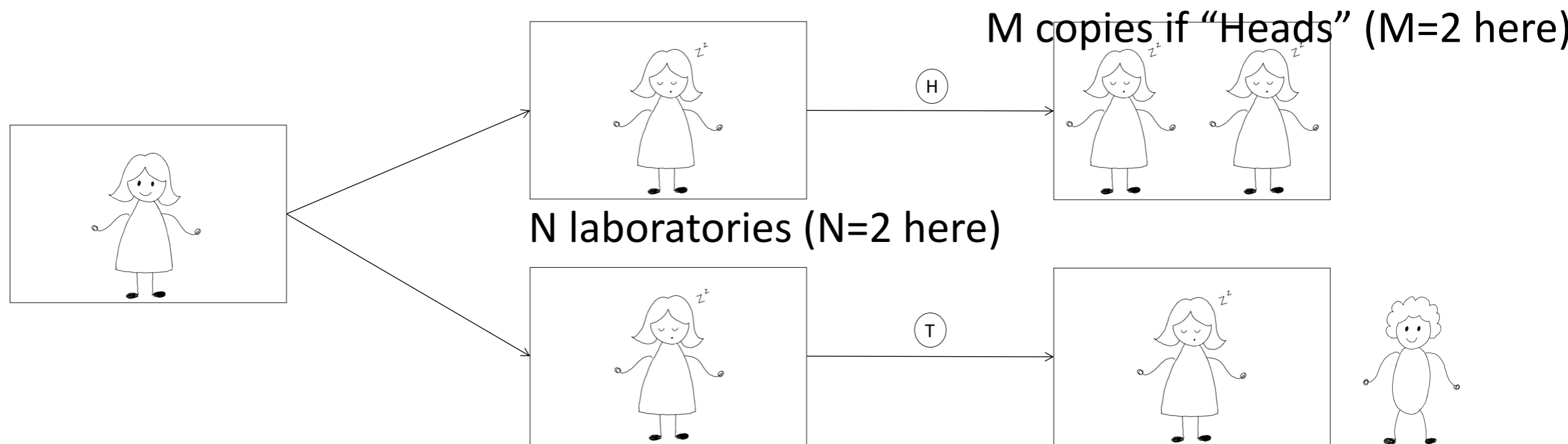


Thought Experiment Imagine Freya is to be put to sleep, and multiplied into N copies. For each copy, now asleep in N different labs, a fair coin is tossed. In each case, if the outcome is Heads, the copy of Freya is duplicated again, or if the outcome is Tails, she is not. Then, each copy of Freya is woken and asked to give her credence that the outcome of her lab's coin toss was Tails. (We assume that she cannot notice the presence/absence of an identical copy of herself in the lab). This scenario is sketched in Figure 3.

In fact, Freya is offered a bet: she can buy a ticket from a bookie for $(2/3 - \varepsilon)\$$, where $\varepsilon > 0$ is small, (say, for 66 cents) that wagers on the coin toss having shown Heads. Meanwhile, superobserving Wigner, outside a certain lab, is offered the same opportunity. It is natural to argue that Freya should buy the ticket, but Wigner should not. That is, the credence that Freya should assign to Tails (which directly determines the maximum price $1 - p$ she rationally ought to be prepared to pay) is $1/3$, whilst for Wigner it is $1/2$.

Finally, all copies of Freya survive the experiment and are released. Everyone who has bought the ticket now receives $1\$$ if the outcome of their lab's coin toss was indeed Tails. Freya and Wigner have been initially informed about all the details of the experiment.

A classical thought experiment



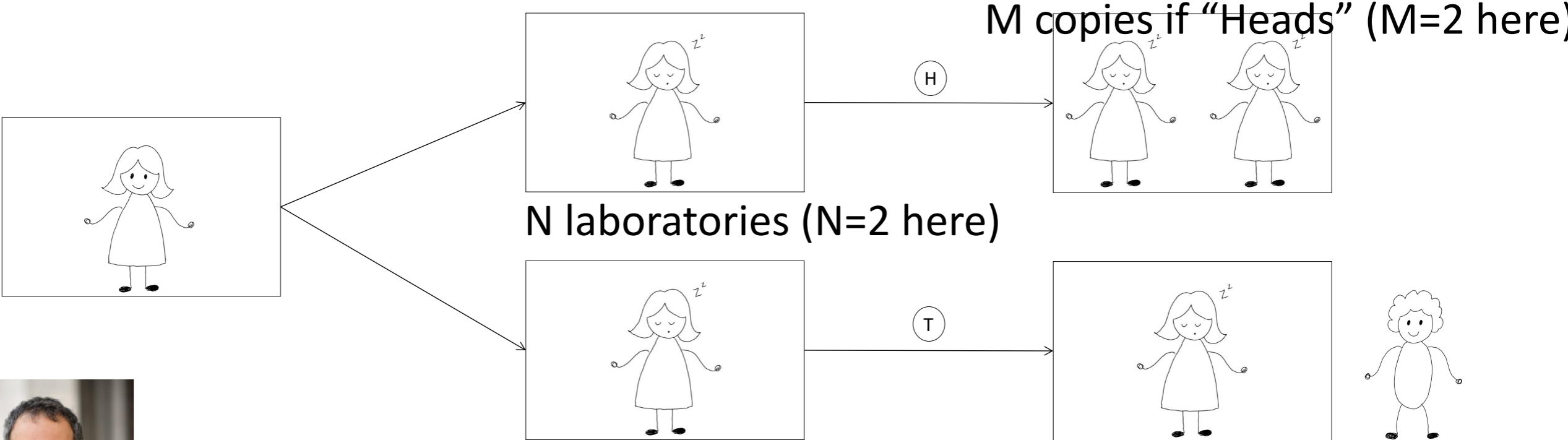
Thought Experiment Imagine Freya is to be put to sleep, and multiplied into N copies. For each copy, now asleep in N different labs, a fair coin is tossed. In each case, if the outcome is Heads, the copy of Freya is duplicated again, or if the outcome is Tails, she is not. Then, each copy of Freya is woken and asked to give her credence that the outcome of her lab's coin toss was Tails. (We assume that she cannot notice the presence/absence of an identical copy of herself in the lab). This scenario is sketched in Figure 3.

In fact, Freya is offered a bet: she can buy a ticket from a bookie for $(2/3 - \varepsilon)\$$, where $\varepsilon > 0$ is small, (say, for 66 cents) that wagers on the coin toss having shown Heads. Meanwhile, superobserving Wigner, outside a certain lab, is offered the same opportunity. It is natural to argue that Freya should buy the ticket, but Wigner should not. That is, the credence that Freya should assign to Tails (which directly determines the maximum price $1 - p$ she rationally ought to be prepared to pay) is $1/3$, whilst for Wigner it is $1/2$.

Finally, all copies of Freya survive the experiment and are released. Everyone who has bought the ticket now receives $1\$$ if the outcome of their lab's coin toss was indeed Tails. Freya and Wigner have been initially informed about all the details of the experiment.

Can make more dramatic ($1/2$ vs. $1/(M+1)$) for M very large.

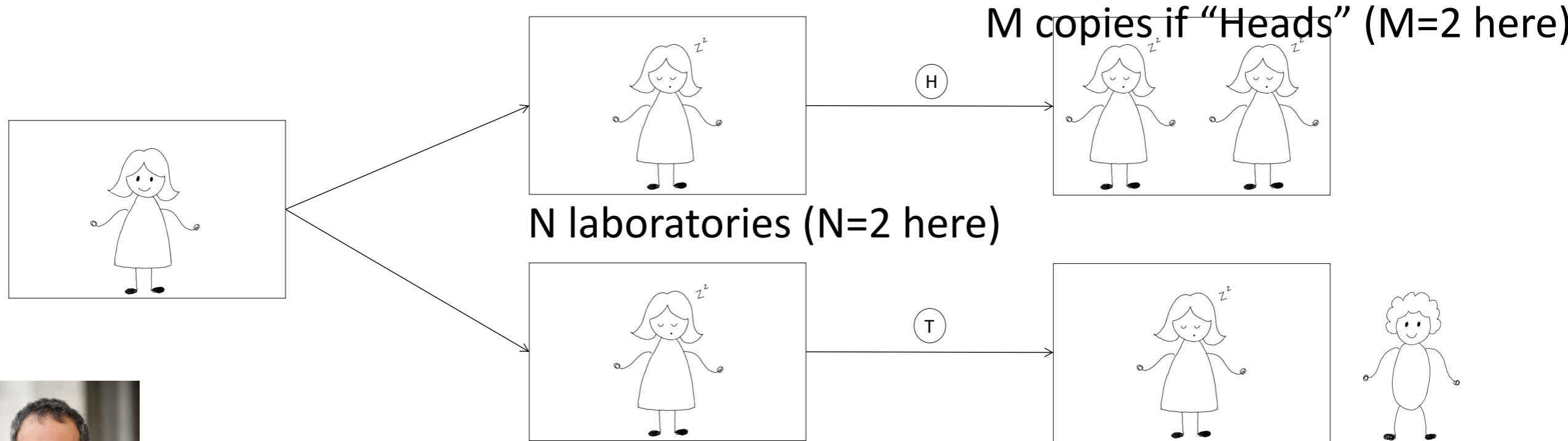
A classical thought experiment



Adam Elga (Princeton)

Elga's Principle of Indifference:
Similar centered worlds deserve equal credence.

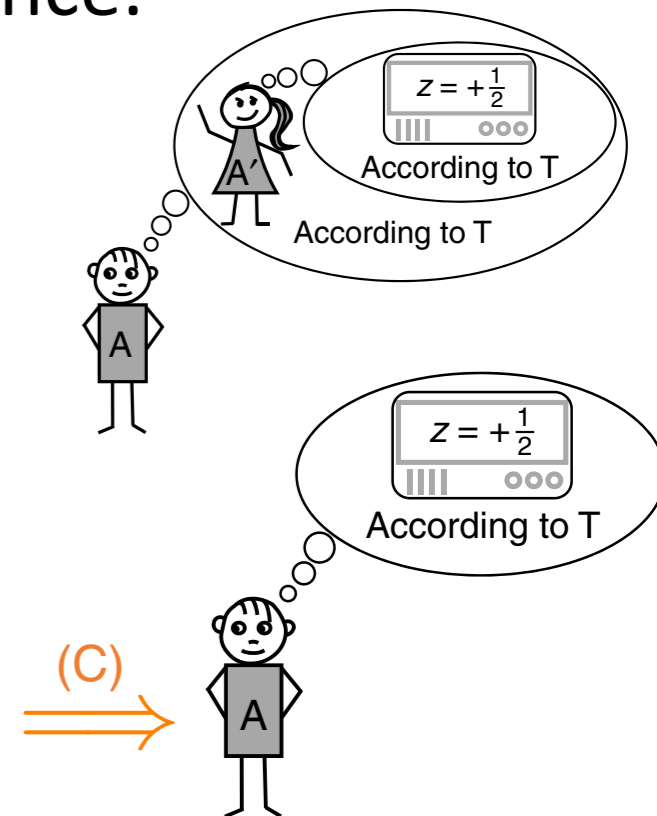
A classical thought experiment



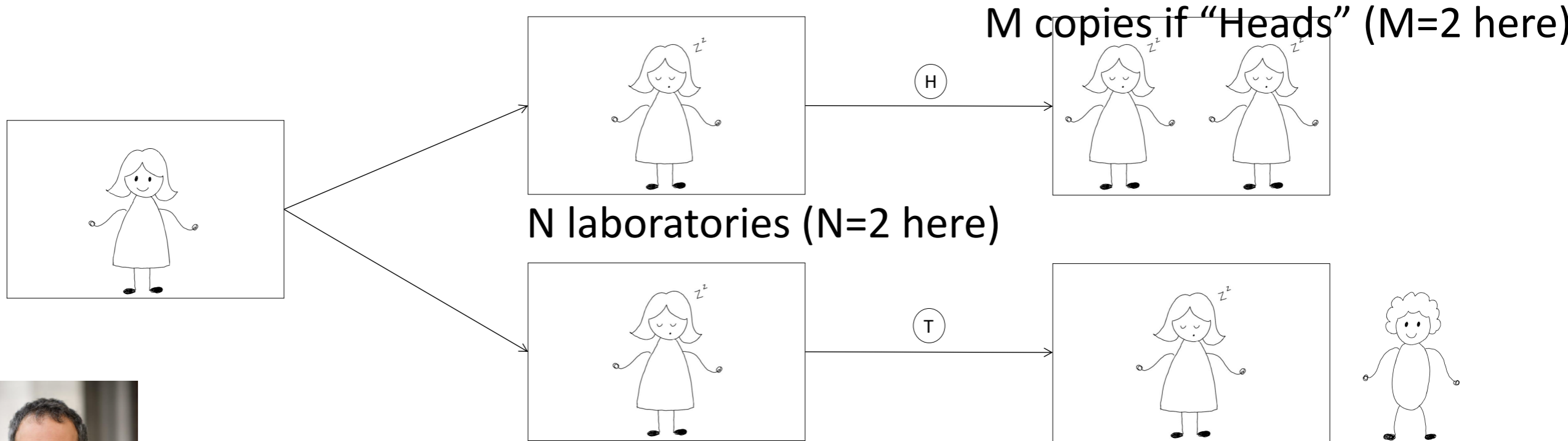
Adam Elga (Princeton)

Elga's Principle of Indifference:
 Similar centered worlds deserve equal credence.

Principle CP (probabilistic consistency): Suppose that an agent A has established that *"I am pretty sure that agent A', upon reasoning within the same theory as the one I am using, and having the exact same knowledge of the world as I, is pretty sure that $x=X$ at time t ".* Then agent A can conclude that *"I am pretty sure that $x=X$ at time t ".*



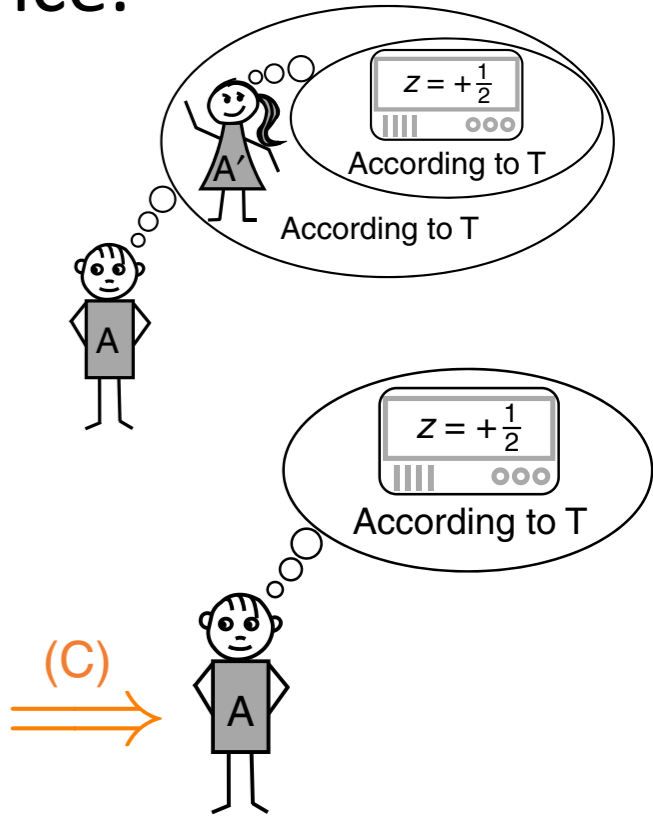
A classical thought experiment



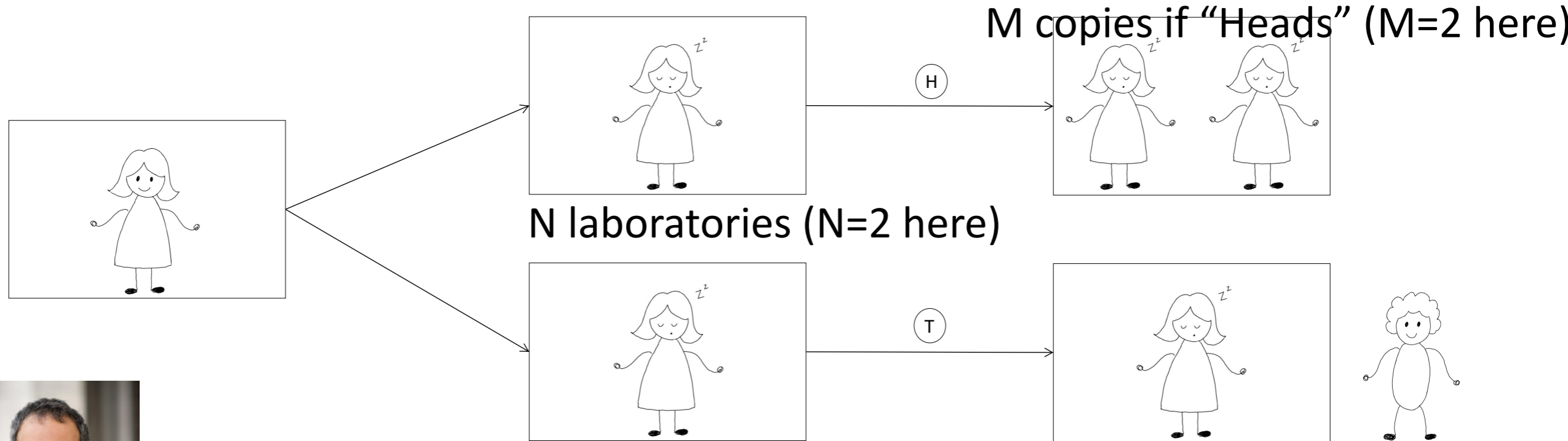
Adam Elga (Princeton)

Elga's Principle of Indifference:
 Similar centered worlds deserve equal credence.

Principle CP (probabilistic consistency)



A classical thought experiment



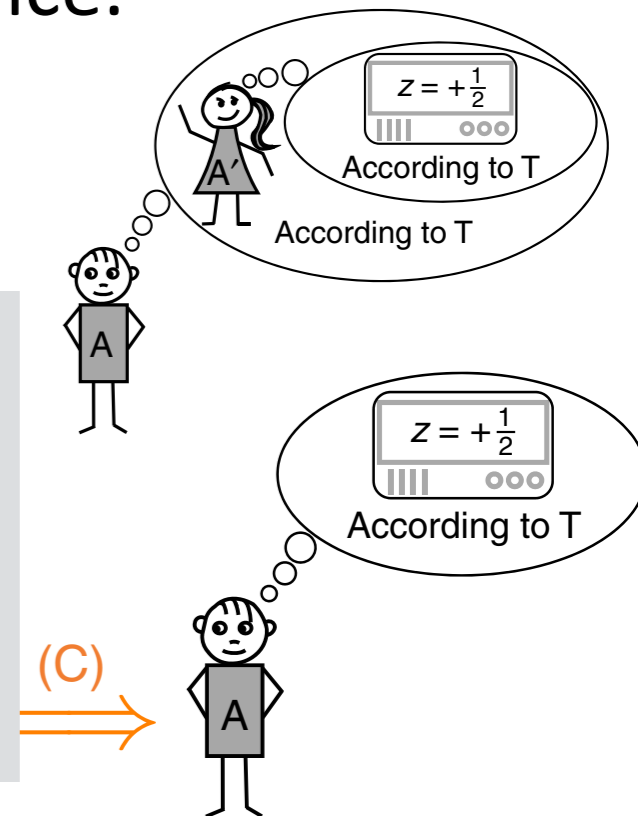
Adam Elga (Princeton)

Elga's Principle of Indifference:

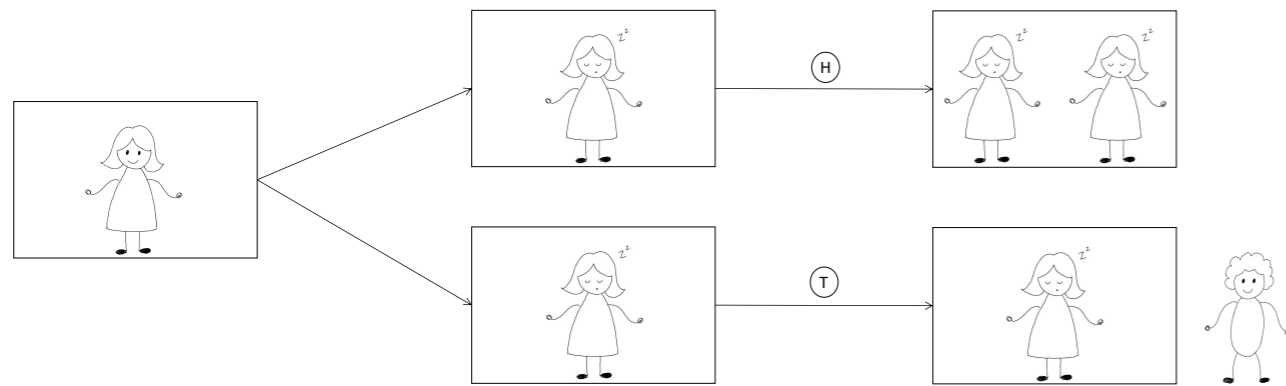
Similar centered worlds deserve equal credence.

Principle CP (probabilistic consistency)

Theorem. Elga's Principle and CP cannot both hold. More generally, as soon as a theory T makes Freya assign probabilities to seeing Heads / Tails that are *not independent of M*, then CP is violated.

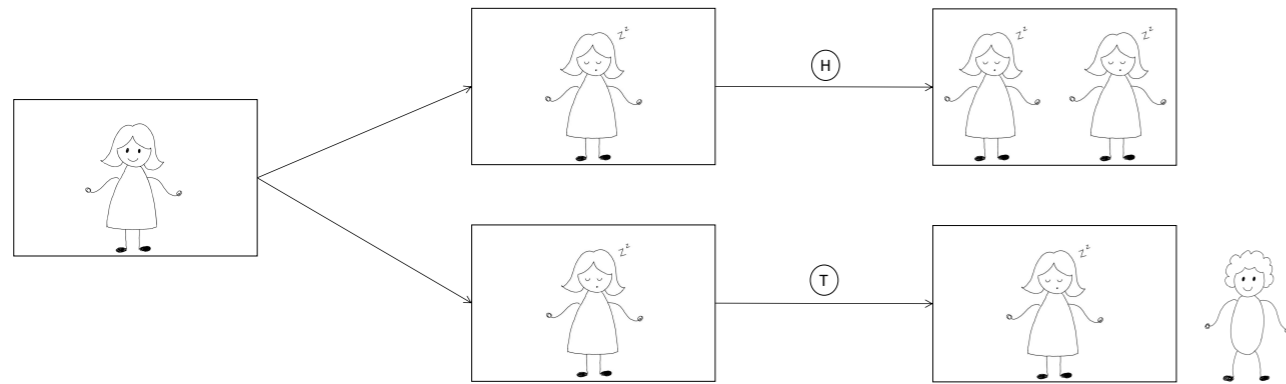


Some clarifications



- We do **not** claim that FR-WF and this duplication scenario are **ontologically** similar!

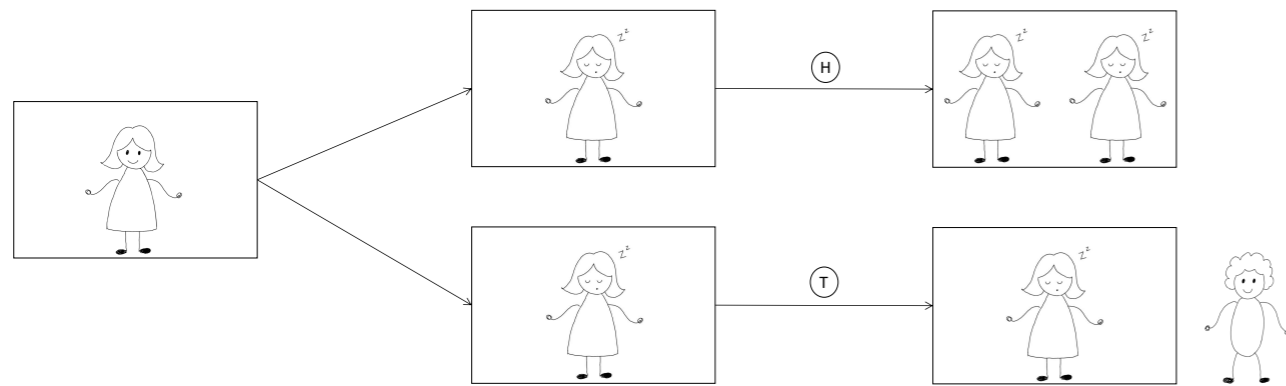
Some clarifications



- We do **not** claim that FR-WF and this duplication scenario are **ontologically** similar!

- **Structural** similarity: there is no joint probability distribution of the observations of Freya and Wigner, if Elga's principle holds (or if Freya's credence depends on M in any other way).

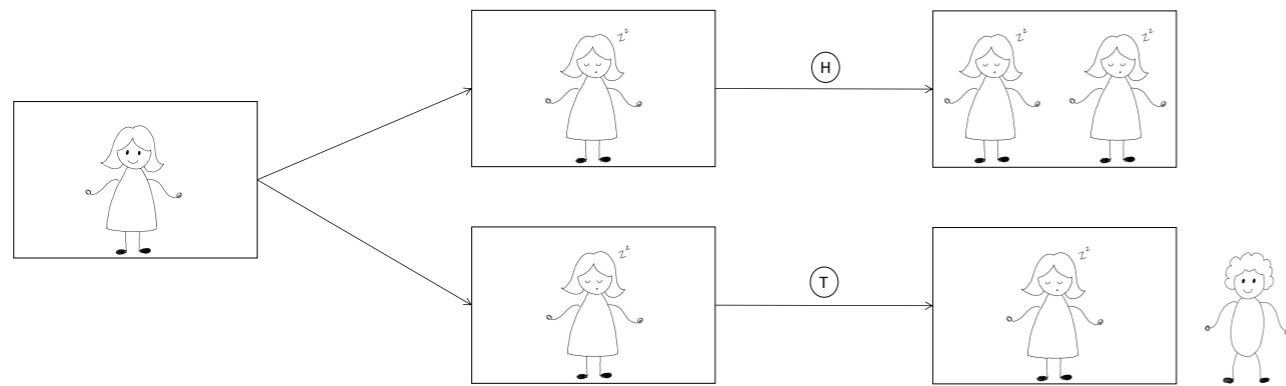
Some clarifications



- We do **not** claim that FR-WF and this duplication scenario are **ontologically** similar!

- **Structural** similarity: there is no joint probability distribution of the observations of Freya and Wigner, if Elga's principle holds (or if Freya's credence depends on M in any other way).
- **Everettians** might see an ontological similarity. But we can turn this around: since branching/dupl. messes up the Kolmogorovian probability space, it admits Everettians to tell their story.

Some clarifications






- We do **not** claim that FR-WF and this duplication scenario are **ontologically** similar!

- **Structural** similarity: there is no joint probability distribution of the observations of Freya and Wigner, if Elga's principle holds (or if Freya's credence depends on M in any other way).
- **Everettians** might see an ontological similarity. But we can turn this around: since branching/dupl. messes up the Kolmogorovian probability space, it admits Everettians to tell their story.
- **Isn't quantum theory natural, but duplication=science fiction?**
No. WF-scenarios are **extremely** invasive on the Friend! Resource requirements for classical duplication are ~ 5 orders of magnitude smaller than Quantum-WF (classical vs. quantum computation).

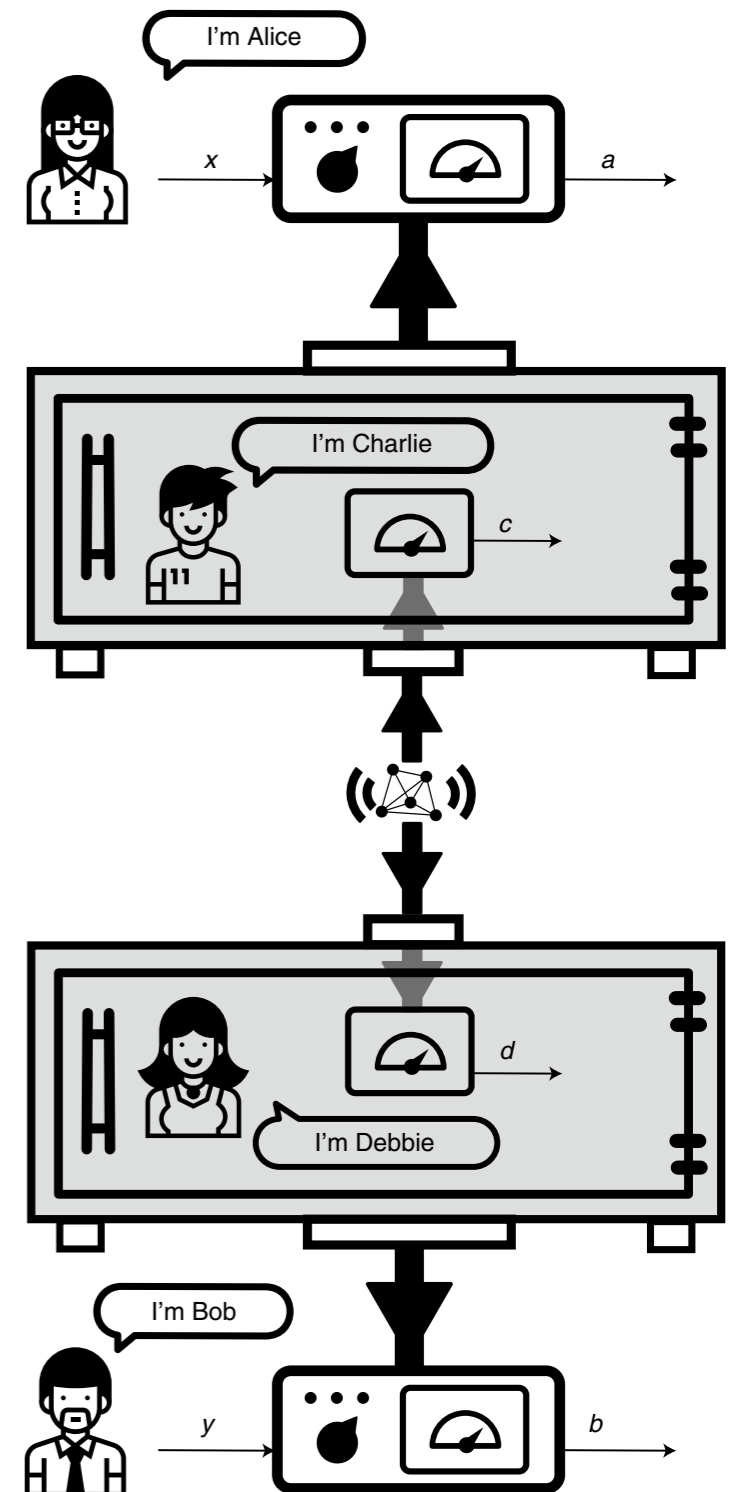


A strong no-go theorem on the Wigner's friend paradox

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

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



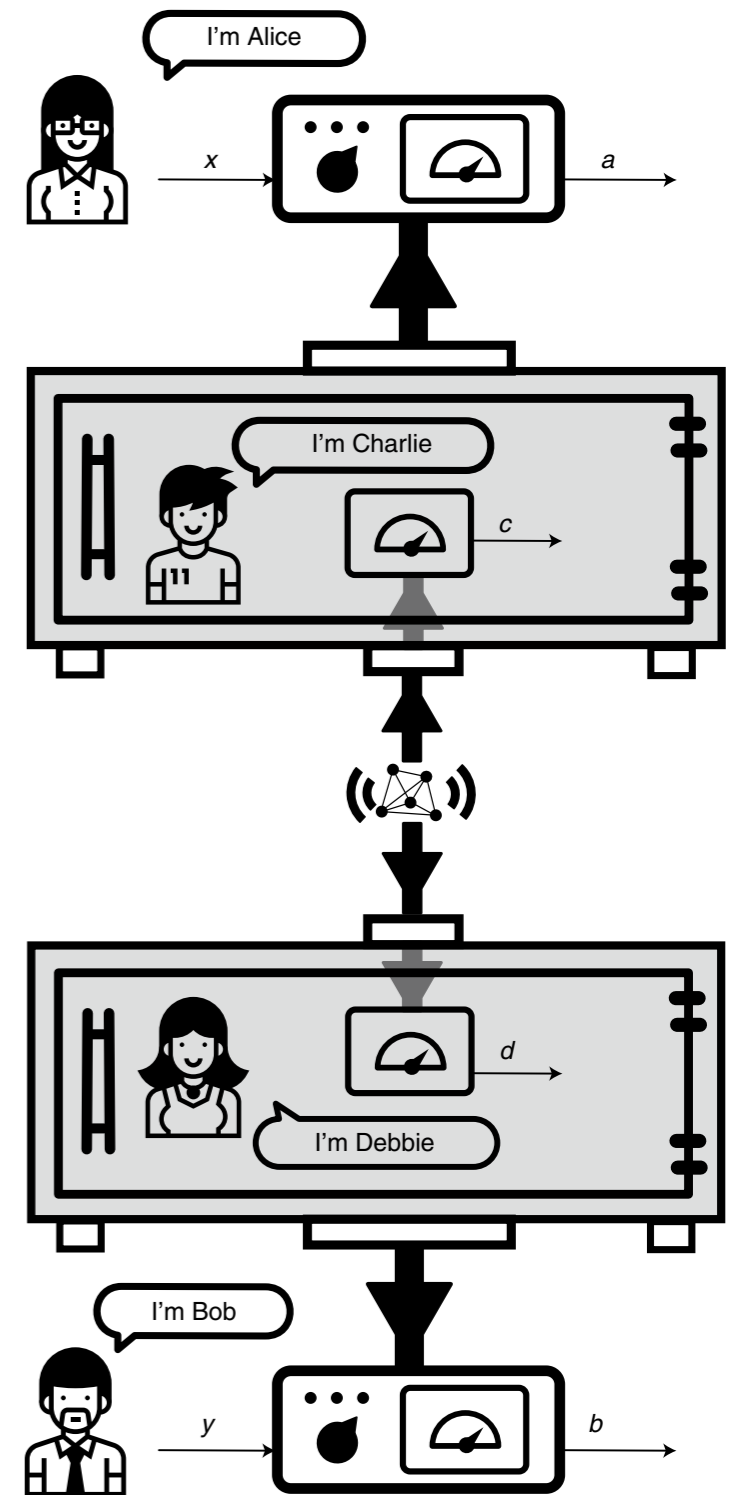
If $x=1$, Alice asks Charlie for its outcome c and outputs $a=c$. (Debbie can be dropped.)



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 $a=c$. (Debbie can be dropped.)

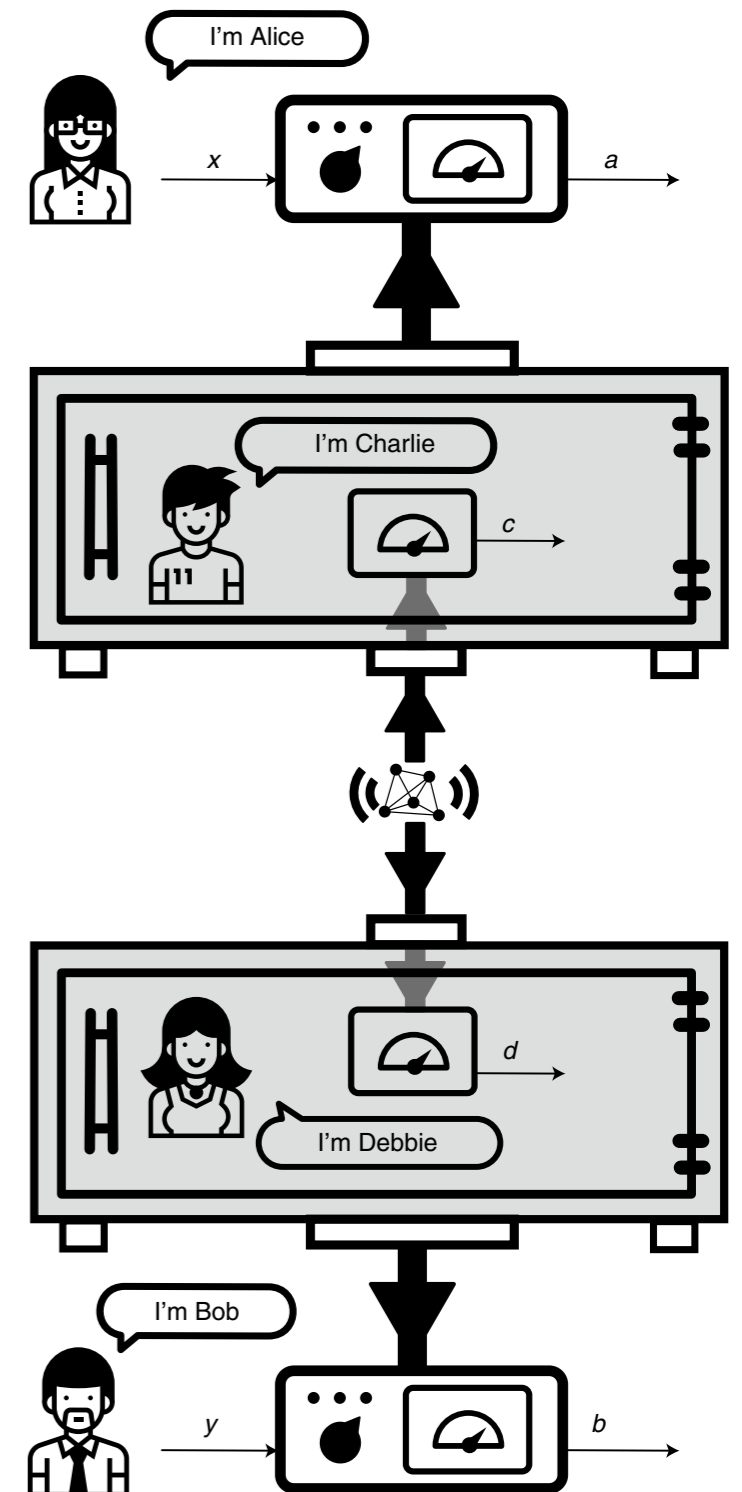
A strong no-go theorem on the Wigner's friend paradox

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:

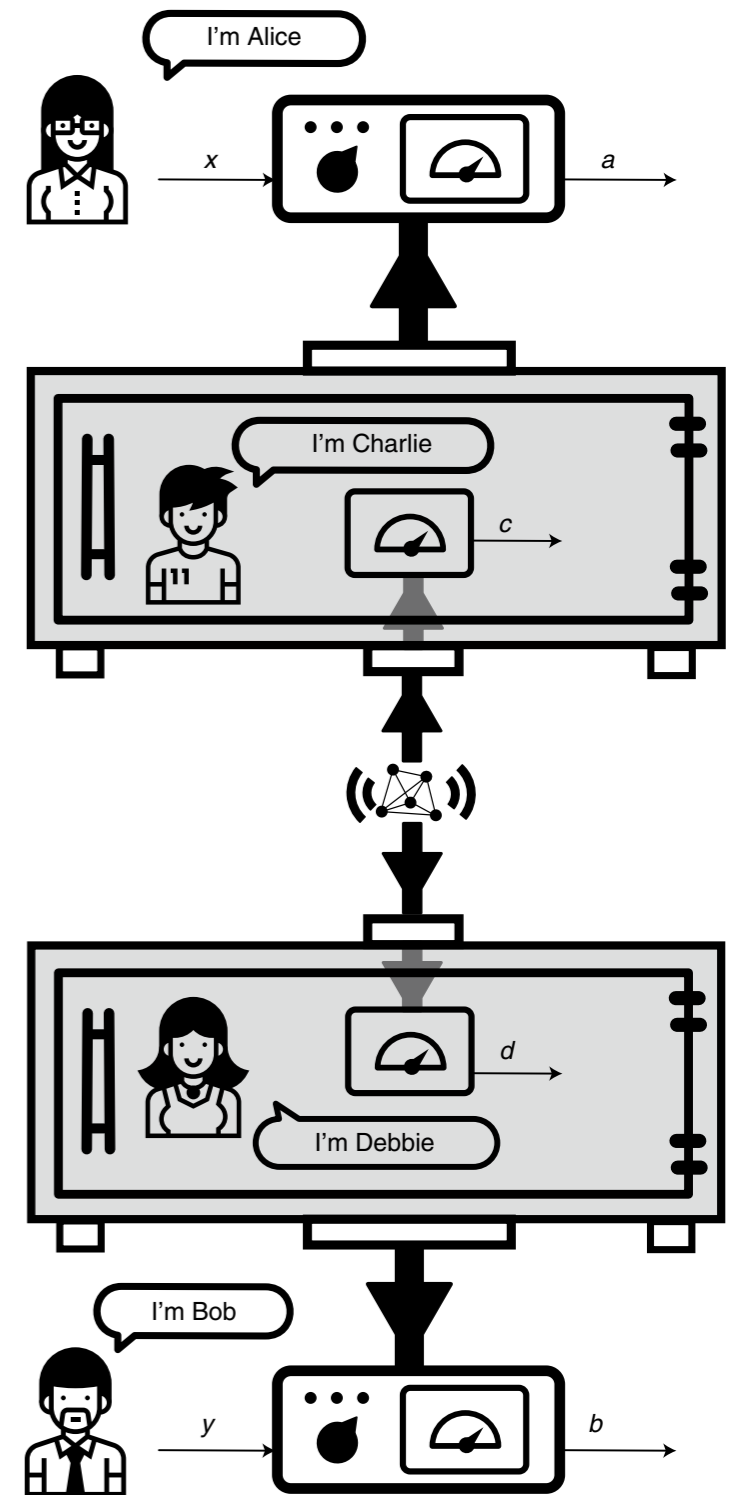
For every x, y , there is a joint prob. distr. $P(a, b, c, d | x, y)$ reproducing the observed distribution $p(a, b | x, y)$ as its marginal.



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

What is (it like to be) Charlie?

From the perspective of A and B (say, using the quantum formalism), **there is no random variable c** , stable over the course of the experiment, describing Charlie's observations.

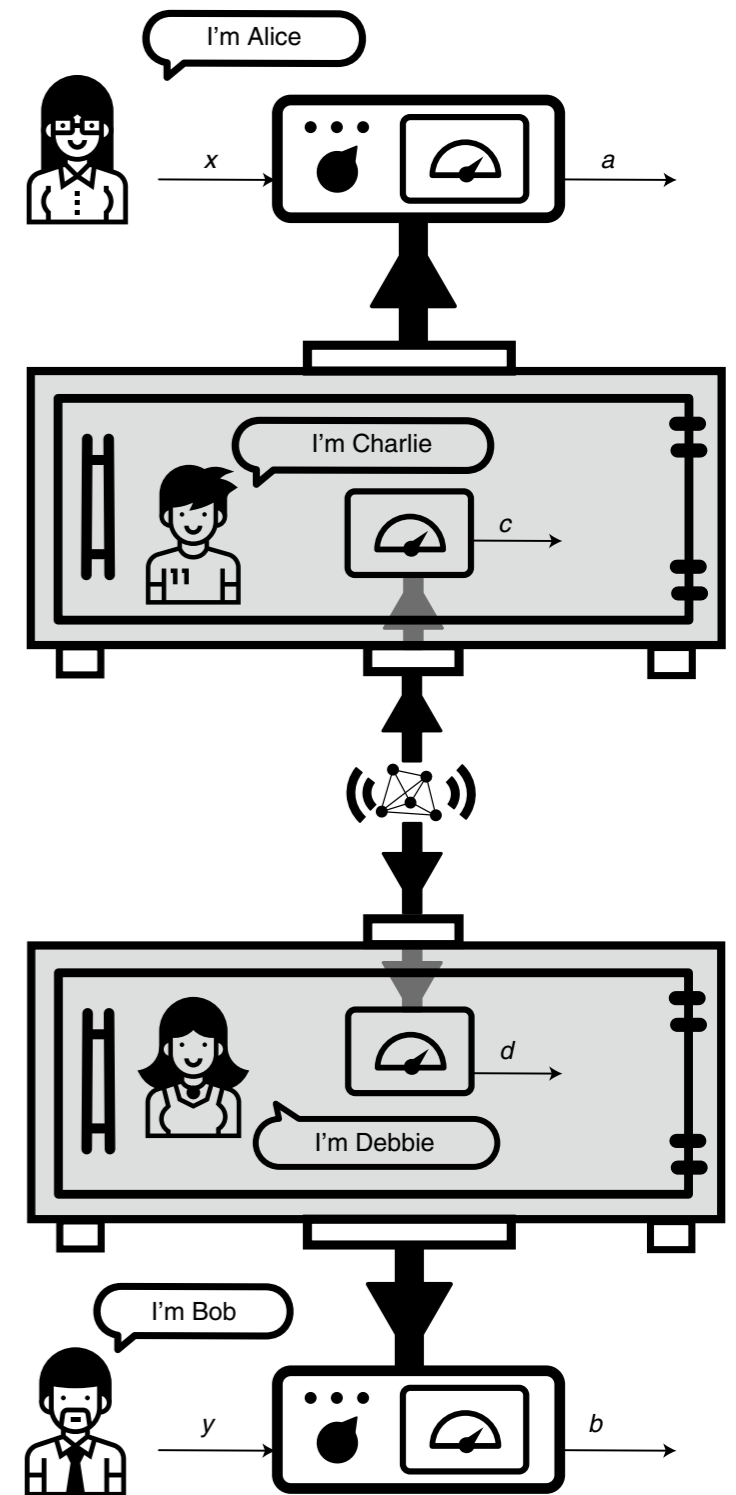


If $x=1$, Alice asks Charlie for its outcome c and outputs $a=c$. (Debbie can be dropped.)

What is (it like to be) Charlie?

From the perspective of A and B (say, using the quantum formalism), **there is no random variable c** , stable over the course of the experiment, describing Charlie's observations.

- **Structural interpretation:** This is ultimately the reason for the non-existence of the joint distributions $P(a,b,c | x,y)$.

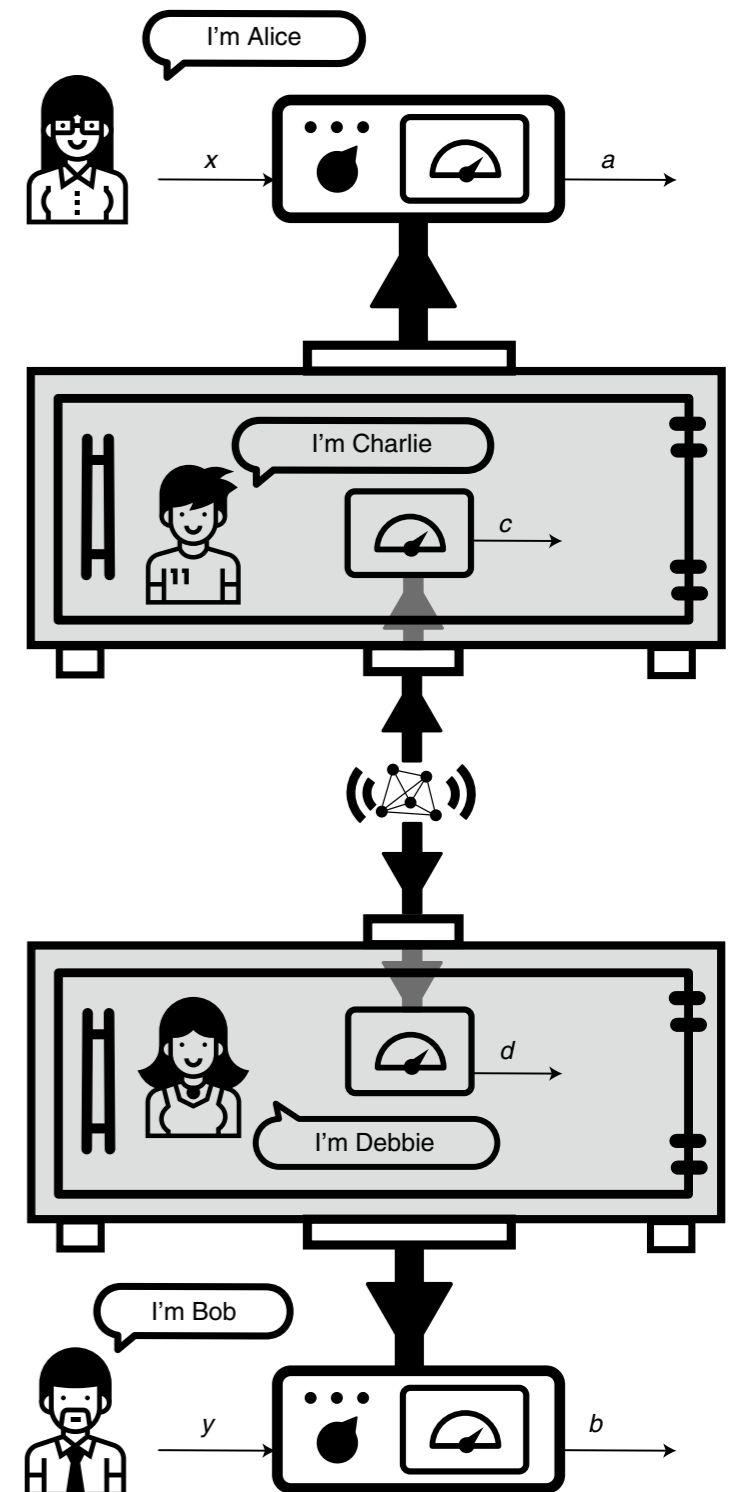


If $x=1$, Alice asks Charlie for its outcome c and outputs $a=c$. (Debbie can be dropped.)

What is (it like to be) Charlie?

From the perspective of A and B (say, using the quantum formalism), **there is no random variable c** , stable over the course of the experiment, describing Charlie's observations.

- **Structural interpretation:** This is ultimately the reason for the non-existence of the joint distributions $P(a,b,c | x,y)$.
- **Conceptual interpretation:** There is no unambiguous external notion of “personal identity” of Charlie over the experiment.



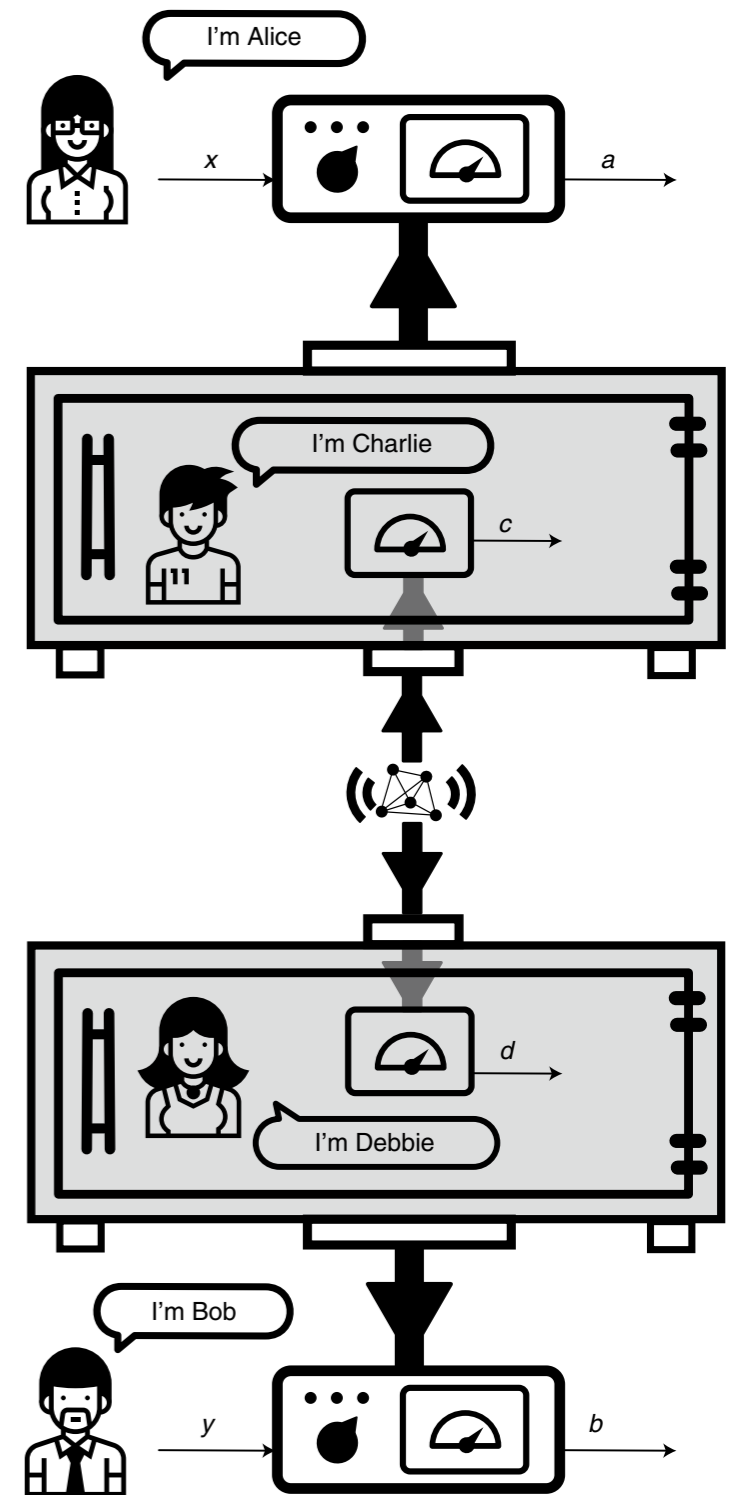
If $x=1$, Alice asks Charlie for its outcome c and outputs $a=c$. (Debbie can be dropped.)

What is (it like to be) Charlie?

From the perspective of A and B (say, using the quantum formalism), **there is no random variable c** , stable over the course of the experiment, describing Charlie's observations.

- **Structural interpretation:** This is ultimately the reason for the non-existence of the joint distributions $P(a,b,c | x,y)$.
- **Conceptual interpretation:** There is no unambiguous external notion of “personal identity” of Charlie over the experiment.

We can simulate this behavior via duplication. To do so, let us look at a reformulation of this WF thought experiment.



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

A “thoughtful” Local Friendliness no-go theorem

H. Wiseman, E. G. Cavalcanti, and E. G. Rieffel, Quantum **7**, 1112 (2023).

A “thoughtful” Local Friendliness no-go theorem

H. Wiseman, E. G. Cavalcanti, and E. G. Rieffel, Quantum **7**, 1112 (2023).

Conceived implementation of the previous thought experiment on a **quantum computer**, proving that the following cannot all hold:

1. **Local Agency:** Any [random] intervention [...] is uncorrelated with any set of physical events that are relevant to that phenomenon and outside the future light-cone of that intervention.
2. **Physical Supervenience:** Any thought supervenes upon some physical process in the brain (or other information-processing unit as appropriate) which can thus be located within a bounded region in space-time.
3. **Ego Absolutism:** My communicable thoughts are absolutely real.
4. **Friendliness:** If [...] an independent party displays cognitive ability at least on par with my own, then they have thoughts, and any thought they communicate is as real as any communicable thought of my own.

A “thoughtful” Local Friendliness no-go theorem

H. Wiseman, E. G. Cavalcanti, and E. G. Rieffel, Quantum **7**, 1112 (2023).

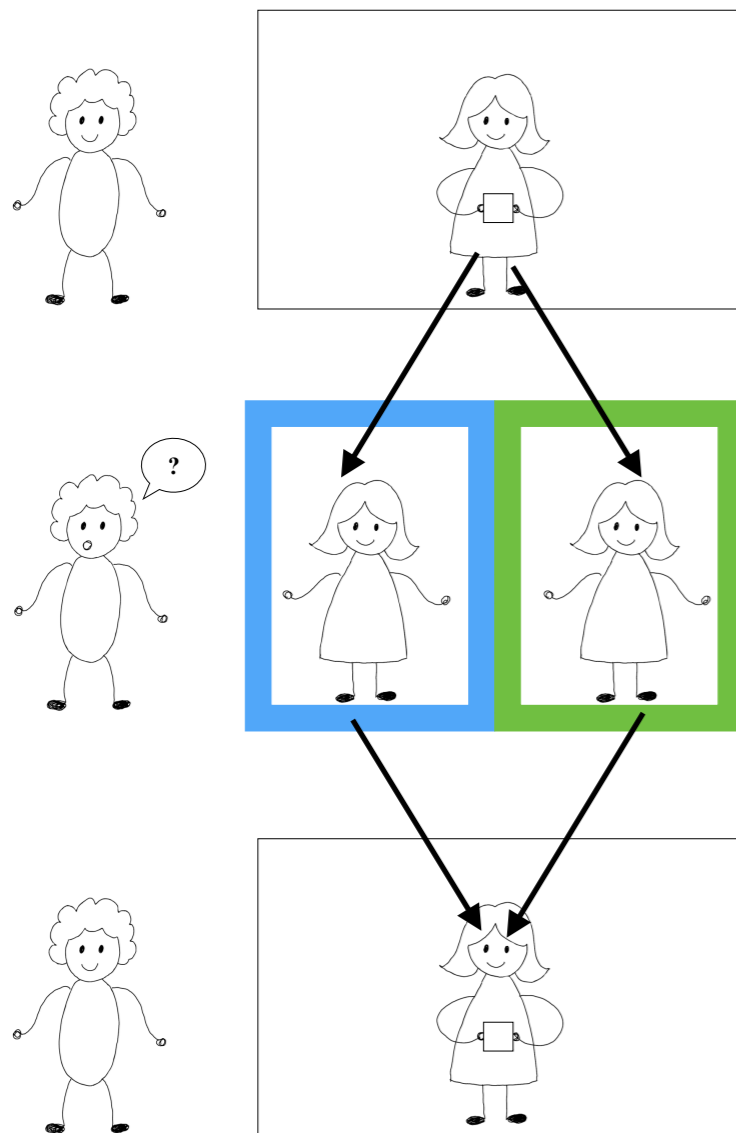
3. **Ego Absolutism:** My communicable thoughts are absolutely real.

A “thoughtful” Local Friendliness no-go theorem

H. Wiseman, E. G. Cavalcanti, and E. G. Rieffel, Quantum **7**, 1112 (2023).

3. **Ego Absolutism:** My communicable thoughts are absolutely real.

Another duplication thought experiment (fission & fusion):

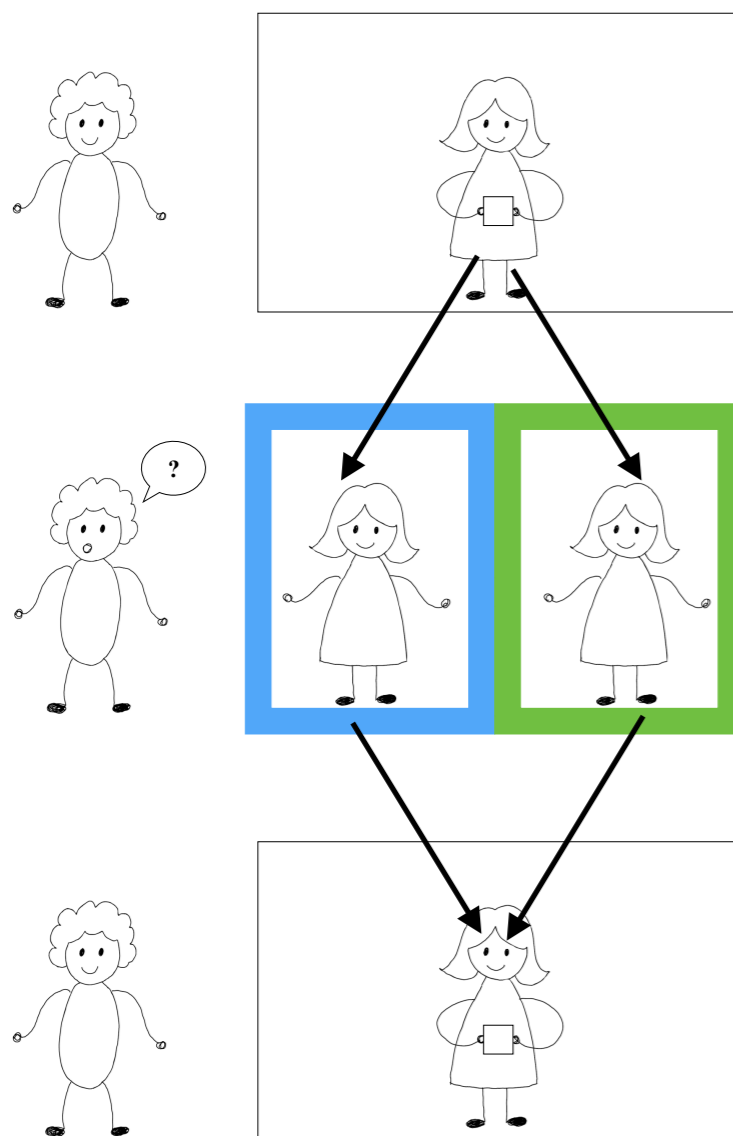


A “thoughtful” Local Friendliness no-go theorem

H. Wiseman, E. G. Cavalcanti, and E. G. Rieffel, Quantum **7**, 1112 (2023).

3. **Ego Absolutism:** My communicable thoughts are absolutely real.

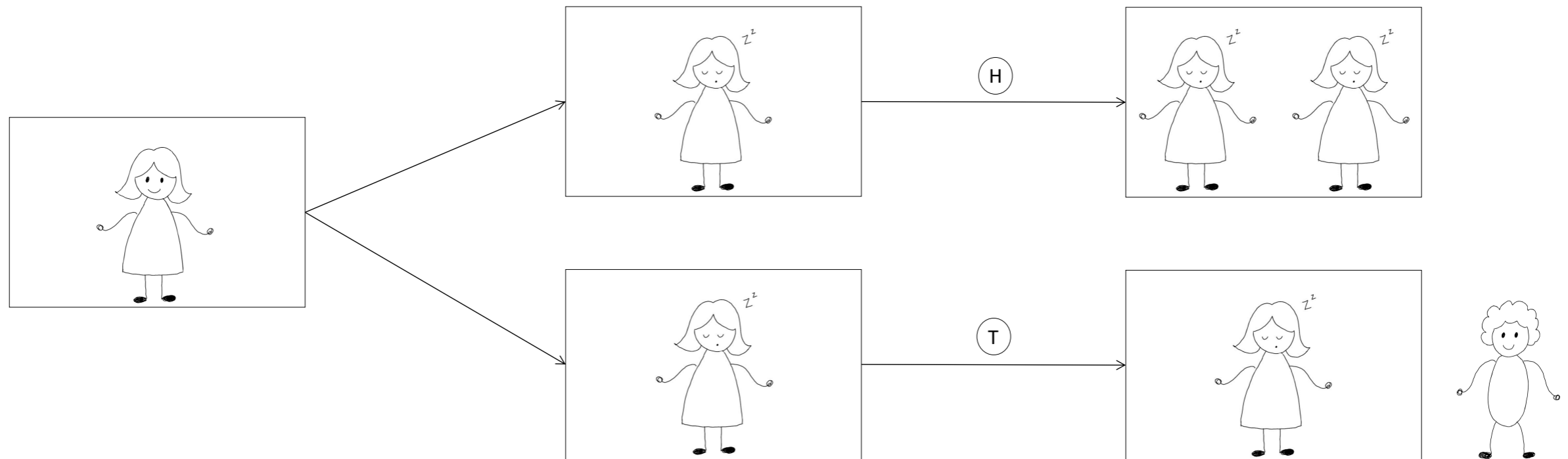
Another duplication thought experiment (fission & fusion):



Naively understood, there exists an ambiguity regarding what “my” indexes for branching scenarios. Returning to the example of Freya, who is yet to be duplicated, she reads Ego Absolutism to say that her communicable thoughts are absolutely real. This includes her thoughts in that instance, such as “I am hungry”. It may also be understood to include thoughts she had this morning, such as “It is raining”. Does it include her future thoughts though? This afternoon, she will be duplicated, whereupon her future copies will have separate experiences. Thus, in describing any future thought she may have, there is an inherent ambiguity as to the meaning of such statements, and whether or not we should take their referent as “absolute”. That is, it is unclear what the words “my (future) thoughts”, if uttered by Freya before the start of the experiment, would refer to, and in disregarding this indexical ambiguity, we will typically be led to mathematical formulations of Ego Absolutism that tacitly involve additional assumptions. In particular, it will lead to the formal assumption that there is always, at every time, a single variable describing a single thought of some person called Freya, while in this branching scenario there are actually two. Indeed, this assumption is part of the mathematical formulation of Bong et al. [7].

Overview

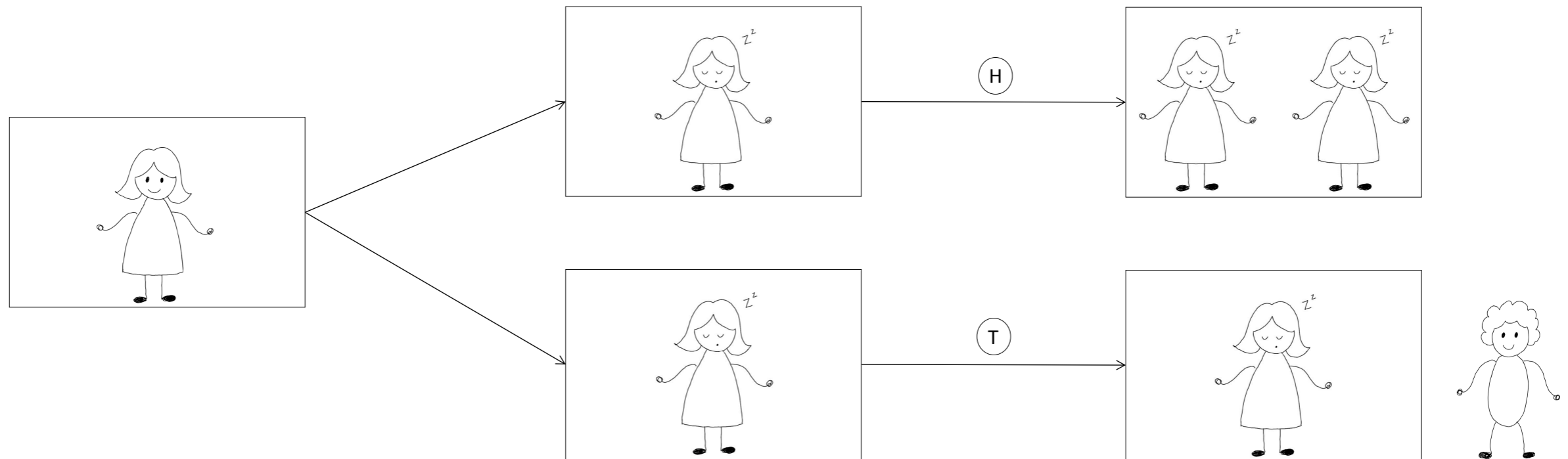
1. Reproducing WF phenomenology with classical **duplication** (“thinking twice inside the box”)



2. A common *structural* core: **Restriction A**
3. Restriction A elsewhere: **Boltzmann brain problem**
4. Consequence: **Fragmentalism/idealism**

Overview

1. Reproducing WF phenomenology with classical **duplication** (“thinking twice inside the box”)



2. A common *structural* core: **Restriction A**

3. Restriction A elsewhere: **Boltzmann brain problem**

4. Consequence: **Fragmentalism/idealism**

A common structural core: Restriction A

A common structural core: Restriction A

We interpret the probabilities in the thought experiments as answers to the question of **what the agent should believe to experience next** — exactly like Philipp Berghofer told us to interpret quantum states yesterday.

A common structural core: Restriction A

We interpret the probabilities in the thought experiments as answers to the question of **what the agent should believe to experience next** — exactly like Philipp Berghofer told us to interpret quantum states yesterday.

Restriction A: Our physical theories do not (sometimes *cannot*) give us joint distributions for the future observations of n **Agents**.

Sometimes even for *single* agents ($n=1$).

This is relative to some theory T and background assumptions.

A common structural core: Restriction A

We interpret the probabilities in the thought experiments as answers to the question of **what the agent should believe to experience next** — exactly like Philipp Berghofer told us to interpret quantum states yesterday.

Restriction A: Our physical theories do not (sometimes *cannot*) give us joint distributions for the future observations of n **Agents**.

Sometimes even for *single* agents ($n=1$).

This is relative to some theory T and background assumptions.

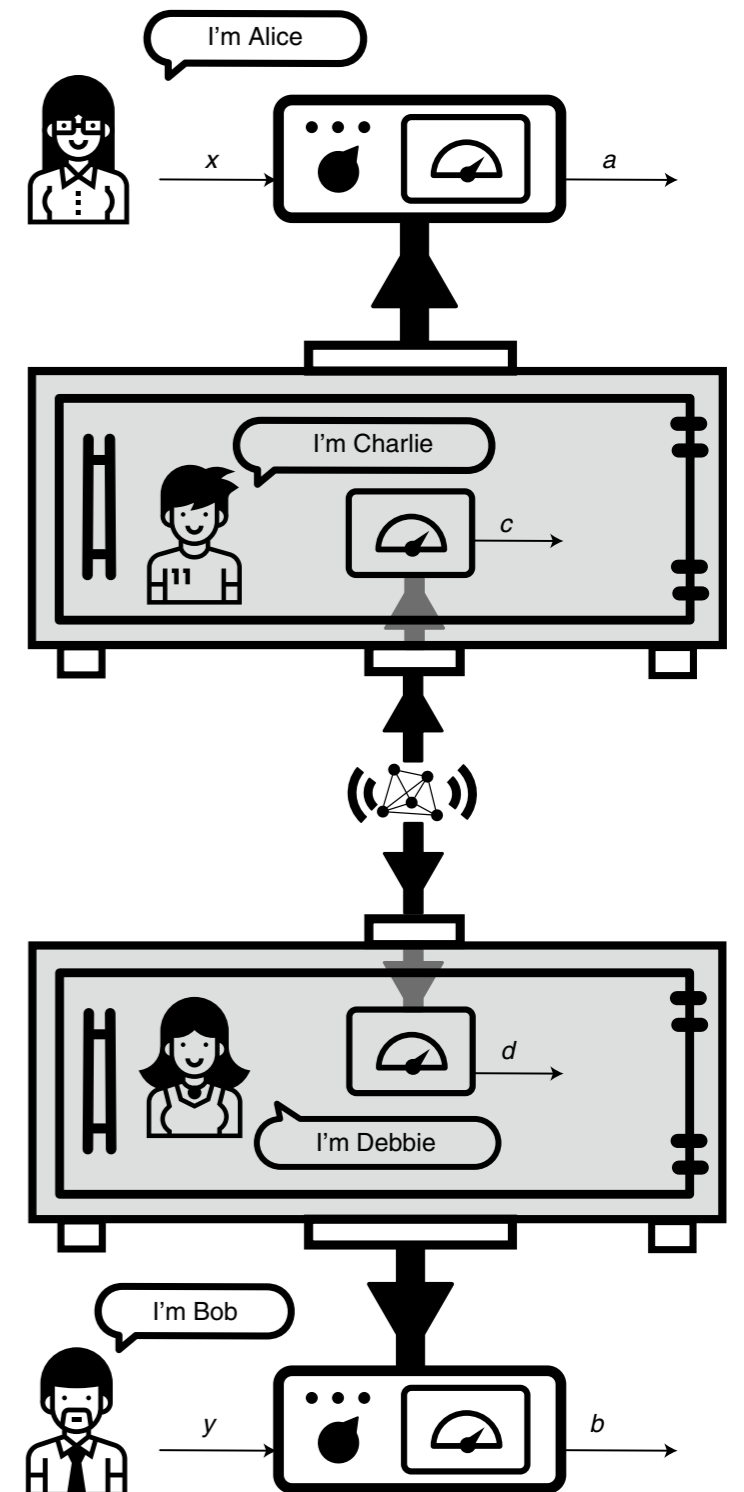
Observation: Unless our physical theory T is empirically incomplete, Restriction A can only apply to situations where it is **impossible** to repeat the scenario identically many times, record the observations of the n agents, and estimate the probabilities via frequencies.

Examples of Restriction A

Examples of Restriction A

- **Bong et al., Absoluteness of Observed Events:** Under the background assumptions of Locality and No Superdeterminism, Restriction A applies to Quantum Theory (in this scenario). Indeed, assuming that an LF-inequality will be experimentally violated, **Restriction A applies to all empirically adequate future physical theories** for $2 \leq n$ agents.

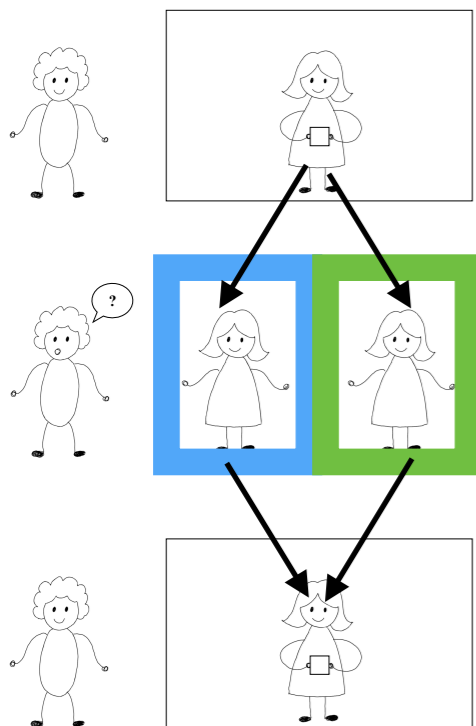
There is no $P(a,b,c|x,y)$.



Examples of Restriction A

- **Bong et al., Absoluteness of Observed Events:** Under the background assumptions of Locality and No Superdeterminism, Restriction A applies to Quantum Theory (in this scenario). Indeed, assuming that an LF-inequality will be experimentally violated, **Restriction A applies to all empirically adequate future physical theories** for $2 \leq n$ agents.

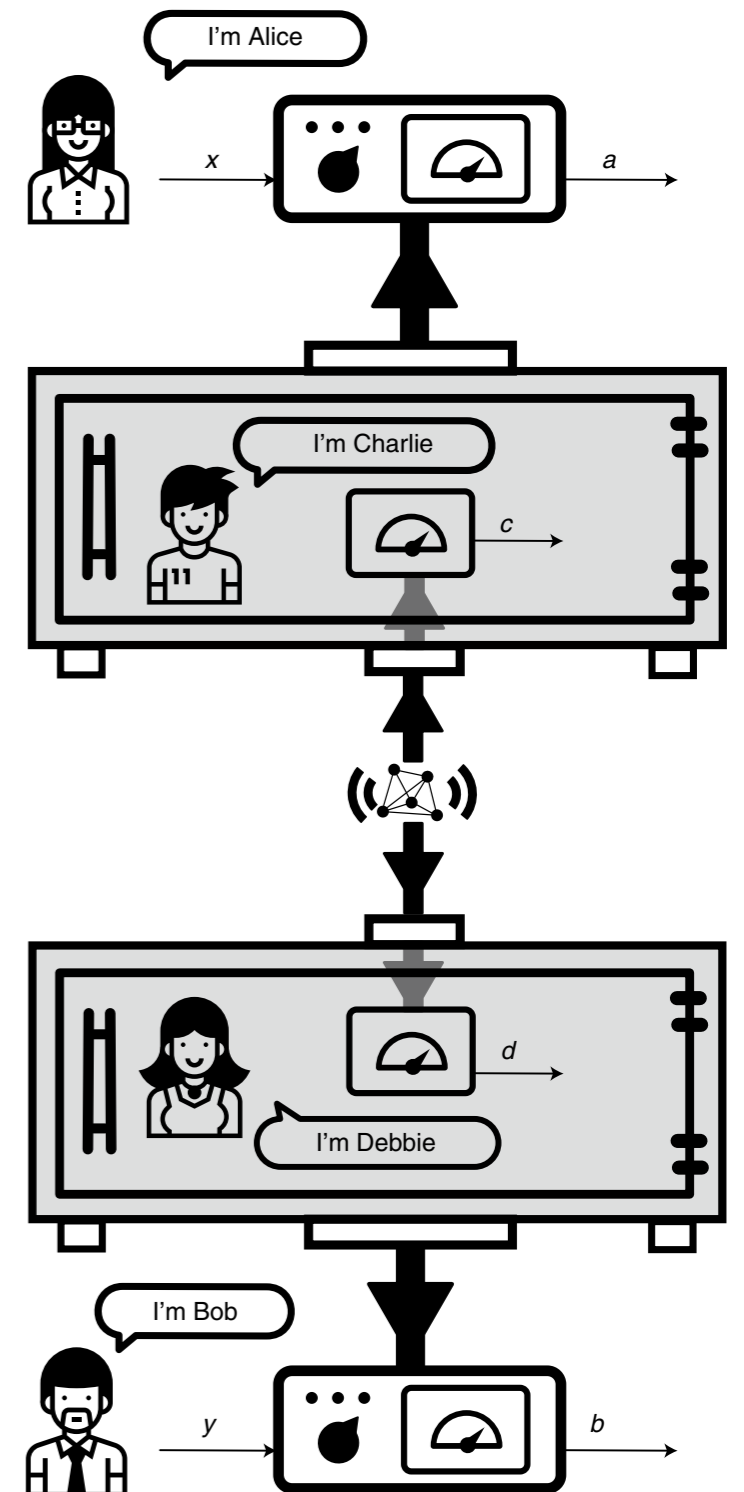
There is no $P(a,b,c|x,y)$.



- **Classical fission&fusion: $n=1$.**

Our physical theories have nothing at all to say about what Freya should believe about the color of room she will see.

Restriction A applies here to all current physical theories.

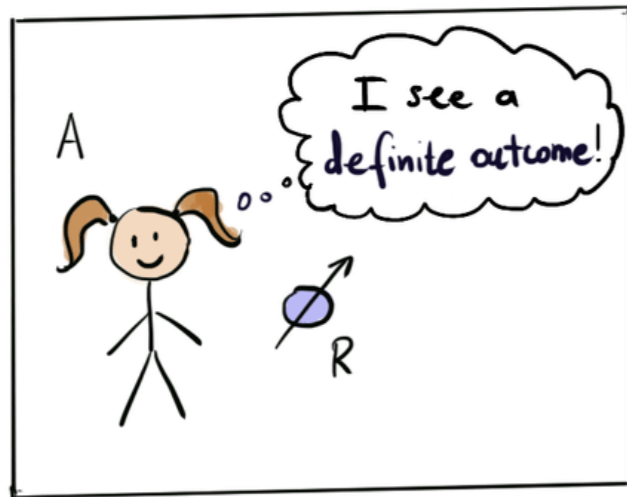


Examples of Restriction A

- **Nonexample: standard Wigner's friend scenario**

Examples of Restriction A

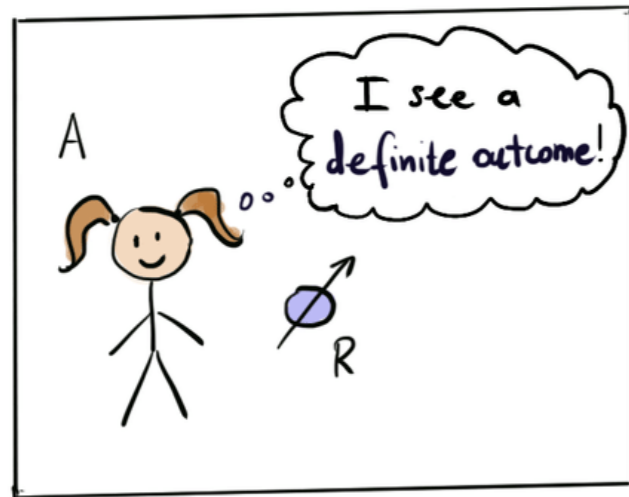
● Nonexample: standard Wigner's friend scenario



Friend measuring the spin, and Wigner in the entangled basis: They can use the Born rule to compute probabilities $P(f)$ and $P(w)$, and the joint distribution is simply the product distribution.

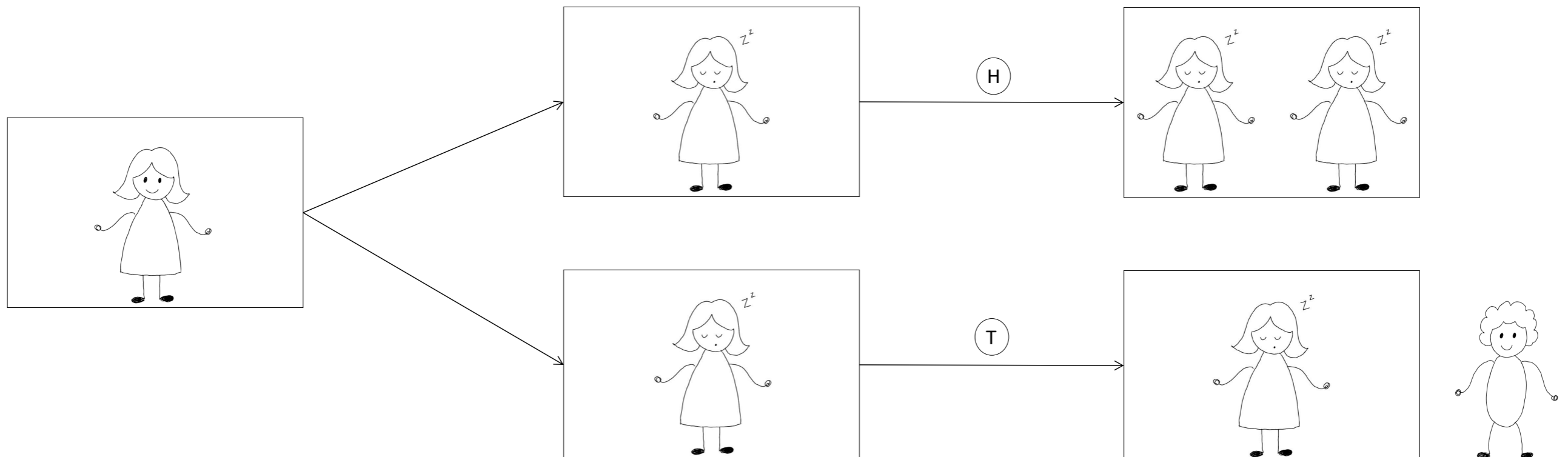
Examples of Restriction A

● Nonexample: standard Wigner's friend scenario



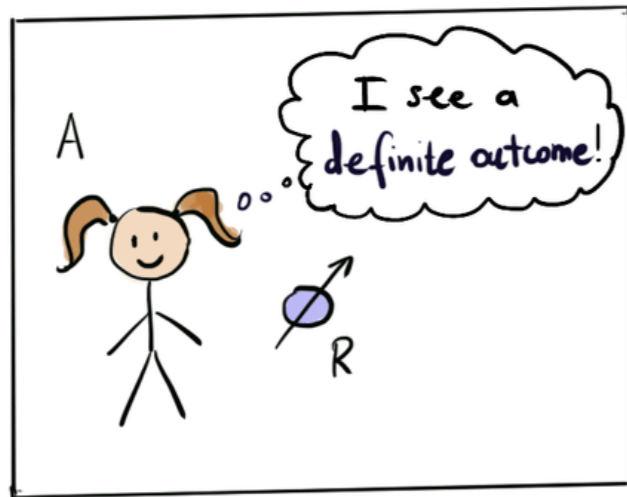
Friend measuring the spin, and Wigner in the entangled basis: They can use the Born rule to compute probabilities $P(f)$ and $P(w)$, and the joint distribution is simply the product distribution.

● Example: violation of probabilistic consistency via classical duplication



Examples of Restriction A

- **Nonexample: standard Wigner's friend scenario**



Friend measuring the spin, and Wigner in the entangled basis: They can use the Born rule to compute probabilities $P(f)$ and $P(w)$, and the joint distribution is simply the product distribution.

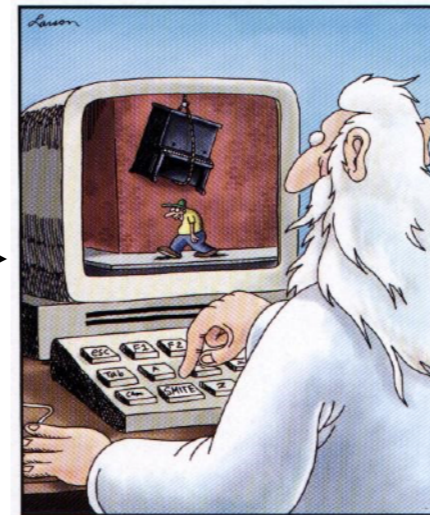
- **Example: violation of probabilistic consistency via classical duplication**

Restriction A applies to this scenario as described by, say, classical physics (because it does not tell us what Freya should believe about her future observations — i.e., to $n=1$ observer). Moreover, even if we supplement classical physics with any probability rule whatsoever (not necessarily Elga's Principle of Indifference), which informs Freya about what she should believe about her future observations *in a way that is not completely ignoring her subsequent multiplicity M* , then the resulting theory will be subject to Restriction A for $2 \leq n$ agents.

Restriction A for $2 \leq n$ agents is unavoidable, for $n=1$ agent **unacceptable**

Restriction A for $2 \leq n$ agents is unavoidable, for $n=1$ agent **unacceptable**

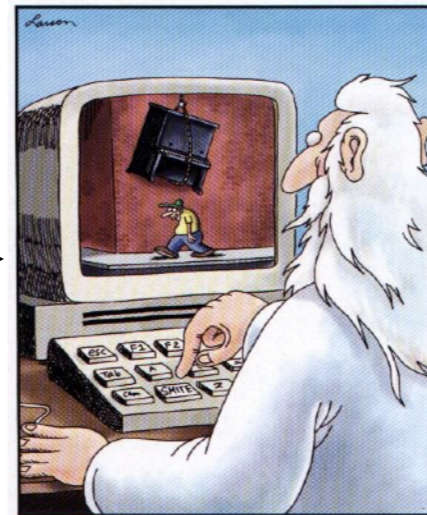
In 2048, you are terminally ill, but the doctor promises to simulate you on a computer when you fall asleep next time (eliminating the original).



God at His computer

Restriction A for $2 \leq n$ agents is unavoidable, for $n=1$ agent **unacceptable**

In 2048, you are terminally ill, but the doctor promises to simulate you on a computer when you fall asleep next time (eliminating the original).



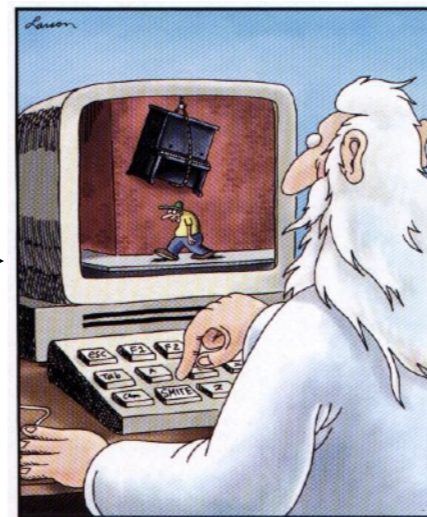
God at His computer

You: Great, but will I *really* wake up in the simulation? Damn, I *really, really want to know!* I'm so afraid! What should I believe will happen to me?

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*?

Restriction A for $2 \leq n$ agents is unavoidable, for $n=1$ agent **unacceptable**

In 2048, you are terminally ill, but the doctor promises to simulate you on a computer when you fall asleep next time (eliminating the original).



God at His computer

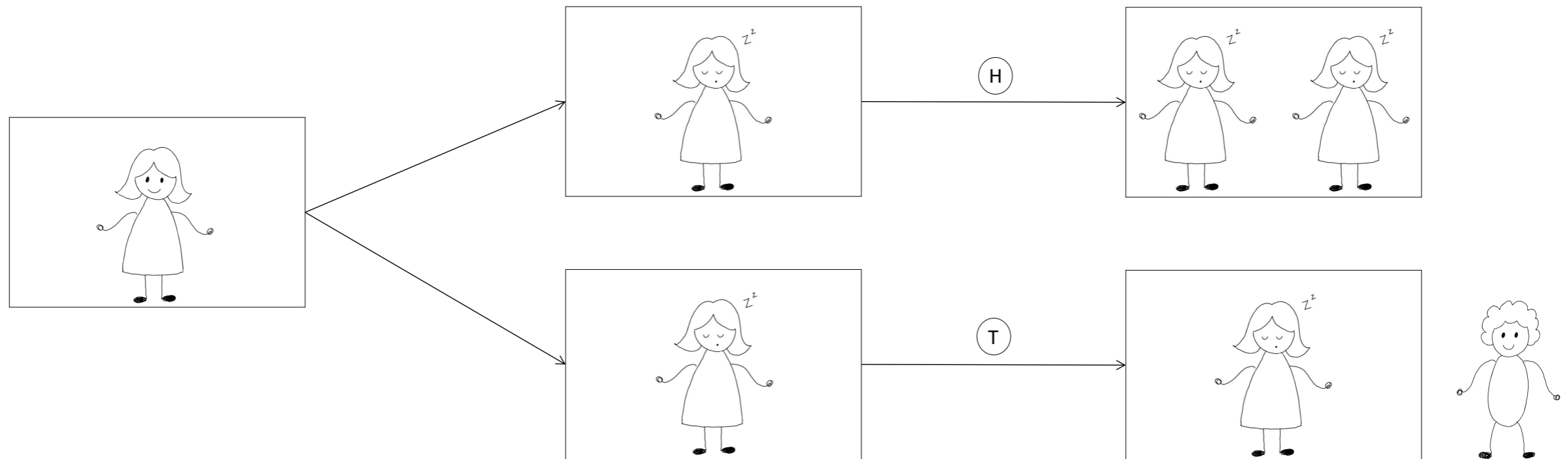
You: Great, but will I *really* wake up in the simulation? Damn, I *really, really want to know!* I'm so afraid! What should I believe will happen to me?

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*?

My claim: This is unacceptable. The first-person perspective is real. There always exists some (degree of) belief that a single agent *should* have — the agent can perform a **private** experiment, and the world will kick back.

Overview

1. Reproducing WF phenomenology with classical **duplication** (“thinking twice inside the box”)



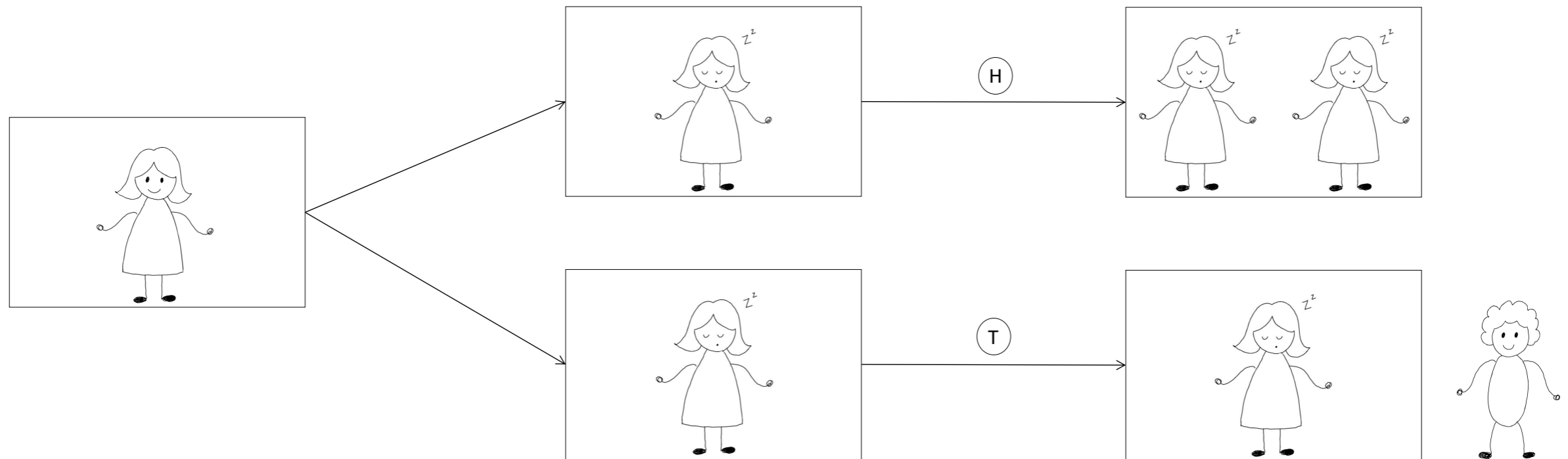
2. A common *structural* core: **Restriction A**

3. Restriction A elsewhere: **Boltzmann brain problem++**

4. Consequence: **Fragmentalism/idealism**

Overview

1. Reproducing WF phenomenology with classical **duplication** (“thinking twice inside the box”)



2. A common *structural* core: **Restriction A**

3. Restriction A elsewhere: **Boltzmann brain problem++**

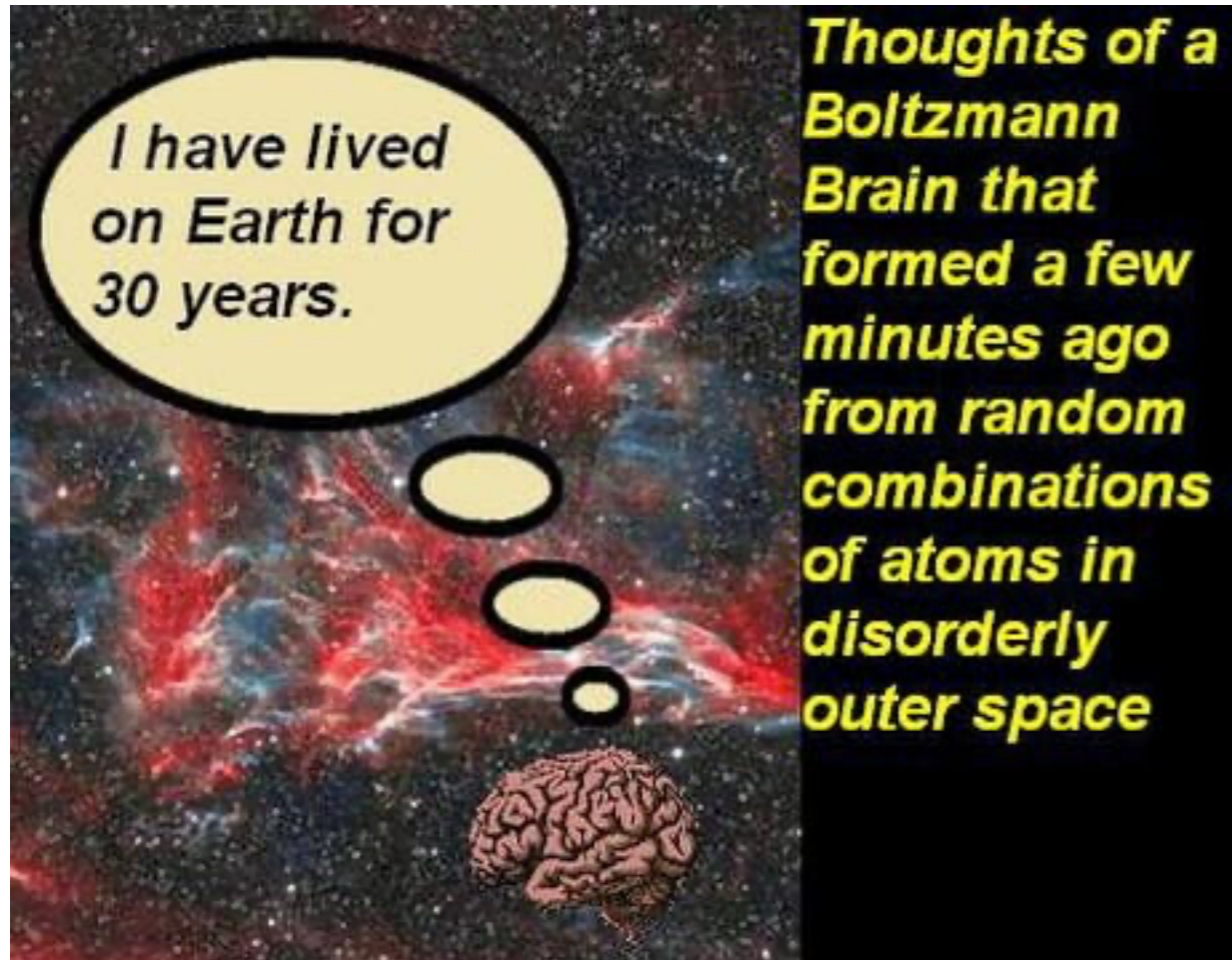
4. Consequence: **Fragmentalism/idealism**

Boltzmann brains and Restriction A

S. M. Carroll, *Why Boltzmann brains are bad*, 2020.

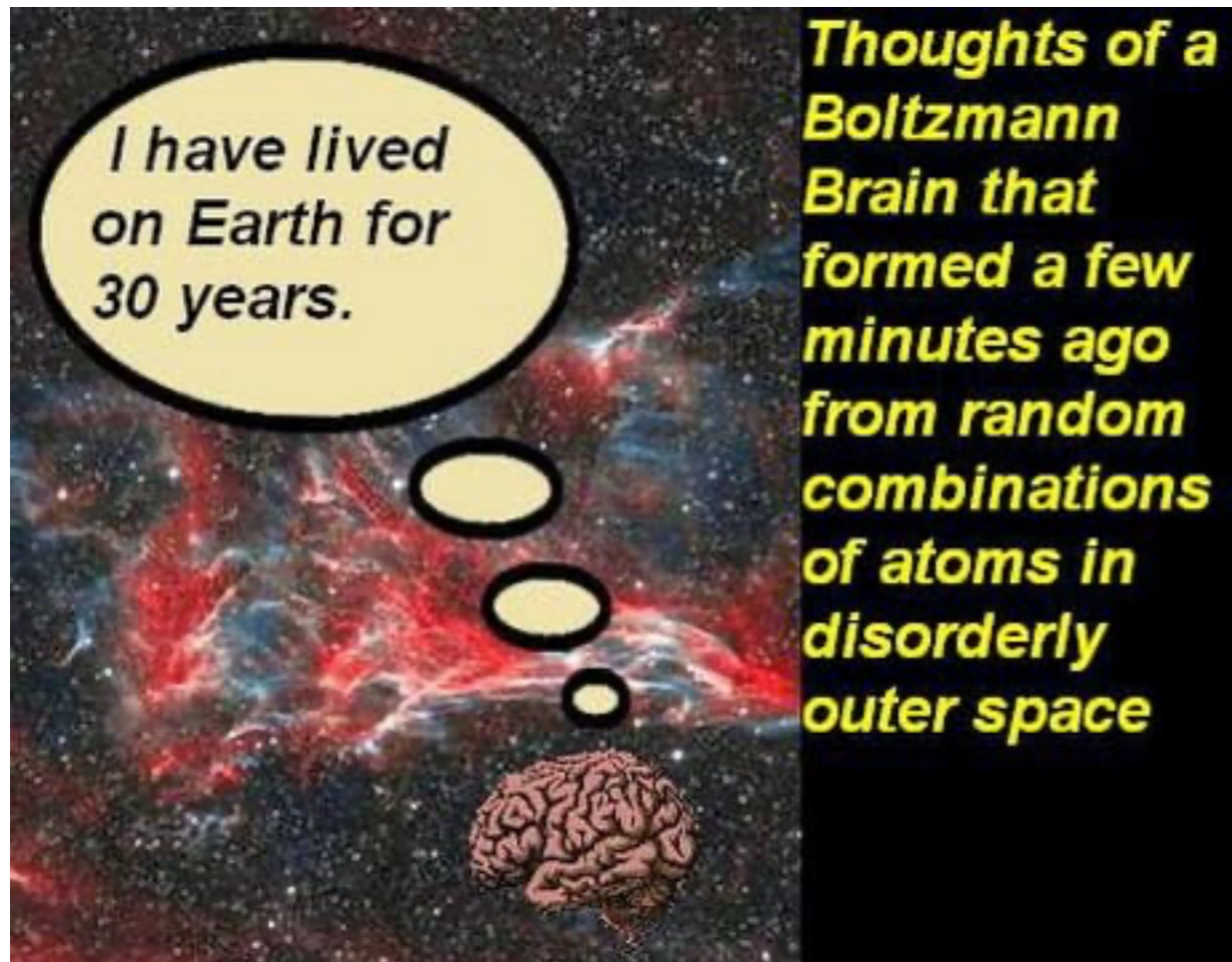
Boltzmann brains and Restriction A

S. M. Carroll, *Why Boltzmann brains are bad*, 2020.



Boltzmann brains and Restriction A

S. M. Carroll, *Why Boltzmann brains are bad*, 2020.

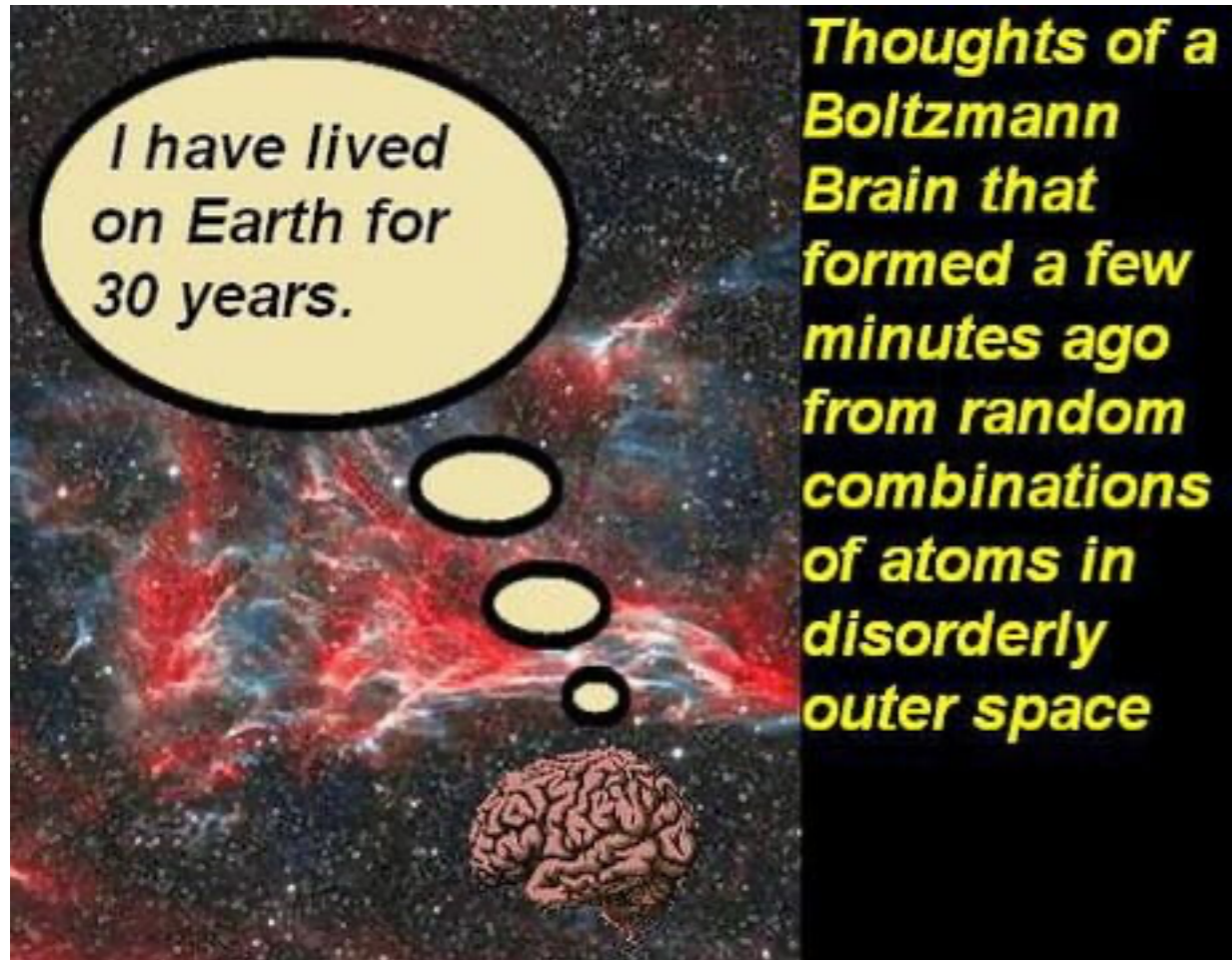


Suppose there is Freya F_0 on Earth, but N copies of Freya F_1, F_2, \dots, F_N out there that came into existence via (something like) thermal fluctuations.

Next, F_0 will make an ordinary observation (such as seeing the microwave background through a telescope), but the other F_i will most likely see something quite unexpected.

Boltzmann brains and Restriction A

S. M. Carroll, *Why Boltzmann brains are bad*, 2020.



Suppose there is Freya F_0 on Earth, but N copies of Freya F_1, F_2, \dots, F_N out there that came into existence via (something like) thermal fluctuations.

Next, F_0 will make an ordinary observation (such as seeing the microwave background through a telescope), but the other F_i will most likely see something quite unexpected.

Question: Is it rational to abandon cosmological models that predict a BB-dominated universe?

Is it rational to abandon BB-dominated models?

Some claim the answer is “yes”, because of either one of the following:

Is it rational to abandon BB-dominated models?

Some claim the answer is “yes”, because of either one of the following:

- (S) The “standard argument”: *“[...] in such a universe, I would probably be a Boltzmann Brain, and I’m not, therefore that’s not the universe in which we live.”* [71]
- (C) Cognitive instability: *“On the one hand, we use our reasoning skills and knowledge of physics to deduce that in such a cosmos we are probably randomly-fluctuated observers, even after conditioning on our local data. On the other hand, we should deduce that we then have no reason to trust those reasoning skills or that knowledge of physics – thus undermining the basis of our argument.”* [71]

Is it rational to abandon BB-dominated models?

Some claim the answer is “yes”, because of either one of the following:

- (S) The “standard argument”: *“[...] in such a universe, I would probably be a Boltzmann Brain, and I’m not, therefore that’s not the universe in which we live.”* [71]
- (C) Cognitive instability: *“On the one hand, we use our reasoning skills and knowledge of physics to deduce that in such a cosmos we are probably randomly-fluctuated observers, even after conditioning on our local data. On the other hand, we should deduce that we then have no reason to trust those reasoning skills or that knowledge of physics – thus undermining the basis of our argument.”* [71]

Both (S) and (C) motivate us to believe that BB-dominated models are false; but both rely on the following crucial assumption:

Is it rational to abandon BB-dominated models?

Some claim the answer is “yes”, because of either one of the following:

- (S) The “standard argument”: *“[...] in such a universe, I would probably be a Boltzmann Brain, and I’m not, therefore that’s not the universe in which we live.”* [71]
- (C) Cognitive instability: *“On the one hand, we use our reasoning skills and knowledge of physics to deduce that in such a cosmos we are probably randomly-fluctuated observers, even after conditioning on our local data. On the other hand, we should deduce that we then have no reason to trust those reasoning skills or that knowledge of physics – thus undermining the basis of our argument.”* [71]

Both (S) and (C) motivate us to believe that BB-dominated models are false; but both rely on the following crucial assumption:

If Freya lives in a BB-dominated universe, she will probably be a BB.

Is it rational to abandon BB-dominated models?

Some claim the answer is “yes”, because of either one of the following:

- (S) The “standard argument”: *“[...] in such a universe, I would probably be a Boltzmann Brain, and I’m not, therefore that’s not the universe in which we live.”* [71]
- (C) Cognitive instability: *“On the one hand, we use our reasoning skills and knowledge of physics to deduce that in such a cosmos we are probably randomly-fluctuated observers, even after conditioning on our local data. On the other hand, we should deduce that we then have no reason to trust those reasoning skills or that knowledge of physics – thus undermining the basis of our argument.”* [71]

Both (S) and (C) motivate us to believe that BB-dominated models are false; but both rely on the following crucial assumption:

If Freya lives in a BB-dominated universe, she will probably be a BB.

It is sufficient to replace this by a weaker and more “empirical” statement:

Is it rational to abandon BB-dominated models?

Some claim the answer is “yes”, because of either one of the following:

- (S) The “standard argument”: *“[...] in such a universe, I would probably be a Boltzmann Brain, and I’m not, therefore that’s not the universe in which we live.”* [71]
- (C) Cognitive instability: *“On the one hand, we use our reasoning skills and knowledge of physics to deduce that in such a cosmos we are probably randomly-fluctuated observers, even after conditioning on our local data. On the other hand, we should deduce that we then have no reason to trust those reasoning skills or that knowledge of physics – thus undermining the basis of our argument.”* [71]

Both (S) and (C) motivate us to believe that BB-dominated models are false; but both rely on the following crucial assumption:

If Freya lives in a BB-dominated universe, she will probably be a BB.

It is sufficient to replace this by a weaker and more “empirical” statement:

If Freya lives in a BB-dominated universe, she should expect to make a weird BB-type observation soon (e.g. seeing infinite-temperature radiation).

Is it rational to abandon BB-dominated models?

Some claim the answer is “yes”, because of either one of the following:

- (S) The “standard argument”: “[...] *in such a universe, I would probably be a Boltzmann Brain, and I’m not, therefore that’s not the universe in which we live.*” [71]
- (C) Cognitive instability: “*On the one hand, we use our reasoning skills and knowledge of physics to deduce that in such a cosmos we are probably randomly-fluctuated observers, even after conditioning on our local data. On the other hand, we should deduce that we then have no reason to trust those reasoning skills or that knowledge of physics – thus undermining the basis of our argument.*” [71]

Both (S) and (C) motivate us to believe that BB-dominated models are false; but both rely on the following crucial assumption:

If Freya lives in a BB-dominated universe, she will probably be a BB.

It is sufficient to replace this by a weaker and more “empirical” statement:

If Freya lives in a BB-dominated universe, she should expect to make a weird BB-type observation soon (e.g. seeing infinite-temperature radiation).

This in turn relies a postulate that relates probabilities with counting:

Is it rational to abandon BB-dominated models?

If there are N copies of Freya in the universe, and M of them are BBs, then Freya should expect with a probability of about M/N to make some strange BB-type observation soon.

Is it rational to abandon BB-dominated models?

If there are N copies of Freya in the universe, and M of them are BBs, then Freya should expect with a probability of about M/N to make some strange BB-type observation soon.

However, this is **not** a statement about facts of the world. It cannot be verified intersubjectively by repeating some experiment many times. Therefore, no contemporary physical theory predicts these (or other) probability assignments for Freya's future observations!

Is it rational to abandon BB-dominated models?

If there are N copies of Freya in the universe, and M of them are BBs, then Freya should expect with a probability of about M/N to make some strange BB-type observation soon.

However, this is **not** a statement about facts of the world. It cannot be verified intersubjectively by repeating some experiment many times. Therefore, no contemporary physical theory predicts these (or other) probability assignments for Freya's future observations!

The claim above is an **additional postulate** (similar to Elga's Principle) that would have to be **added** to our physical theories. It is a particular way to react to Restriction A, but by far not the only (natural) one.

Is it rational to abandon BB-dominated models?

If there are N copies of Freya in the universe, and M of them are BBs, then Freya should expect with a probability of about M/N to make some strange BB-type observation soon.

However, this is **not** a statement about facts of the world. It cannot be verified intersubjectively by repeating some experiment many times. Therefore, no contemporary physical theory predicts these (or other) probability assignments for Freya's future observations!

The claim above is an **additional postulate** (similar to Elga's Principle) that would have to be **added** to our physical theories. It is a particular way to react to Restriction A, but by far not the only (natural) one.

Conclusion: Restriction A applies, and cosmologists **cannot** simply abandon BB-dominated cosmological models.

Is it rational to abandon BB-dominated models?

If there are N copies of Freya in the universe, and M of them are BBs, then Freya should expect with a probability of about M/N to make some strange BB-type observation soon.

However, this is **not** a statement about facts of the world. It cannot be verified intersubjectively by repeating some experiment many times. Therefore, no contemporary physical theory predicts these (or other) probability assignments for Freya's future observations!

The claim above is an **additional postulate** (similar to Elga's Principle) that would have to be **added** to our physical theories. It is a particular way to react to Restriction A, but by far not the only (natural) one.

Conclusion: Restriction A applies, and cosmologists **cannot** simply abandon BB-dominated cosmological models.

After all, the above is not *that* plausible: how to count? What if $M=N=\infty$?

Is it rational to abandon BB-dominated models?

If there are N copies of Freya in the universe, and M of them are BBs, then Freya should expect with a probability of about M/N to make some strange BB-type observation soon.

Is it rational to abandon BB-dominated models?

If there are N copies of Freya in the universe, and M of them are BBs, then Freya should expect with a probability of about M/N to make some strange BB-type observation soon.

Suggestion to replace it:

Future observations are more likely if **universal induction** tells us so (*not* if they are more multiply realized). Essentially, use an algorithmic prior to predict future observations.

This would imply that Freya will probably continue seeing “Earth-like business as usual” even if Earth-Freya was hugely outnumbered by BBs.

Is it rational to abandon BB-dominated models?

If there are N copies of Freya in the universe, and M of them are BBs, then Freya should expect with a probability of about M/N to make some strange BB-type observation soon.

Suggestion to replace it:

Future observations are more likely if **universal induction** tells us so (*not* if they are more multiply realized). Essentially, use an algorithmic prior to predict future observations.

This would imply that Freya will probably continue seeing “Earth-like business as usual” even if Earth-Freya was hugely outnumbered by BBs.

To make sense of this statement, the agent “Freya” would not be defined by a localized physical system on which she supervenes, but by her **structural properties**, represented both on Earth and in the BBs.

Is it rational to abandon BB-dominated models?

If there are N copies of Freya in the universe, and M of them are BBs, then Freya should expect with a probability of about M/N to make some strange BB-type observation soon.

Suggestion to replace it:

Future observations are more likely if **universal induction** tells us so (*not* if they are more multiply realized). Essentially, use an algorithmic prior to predict future observations.

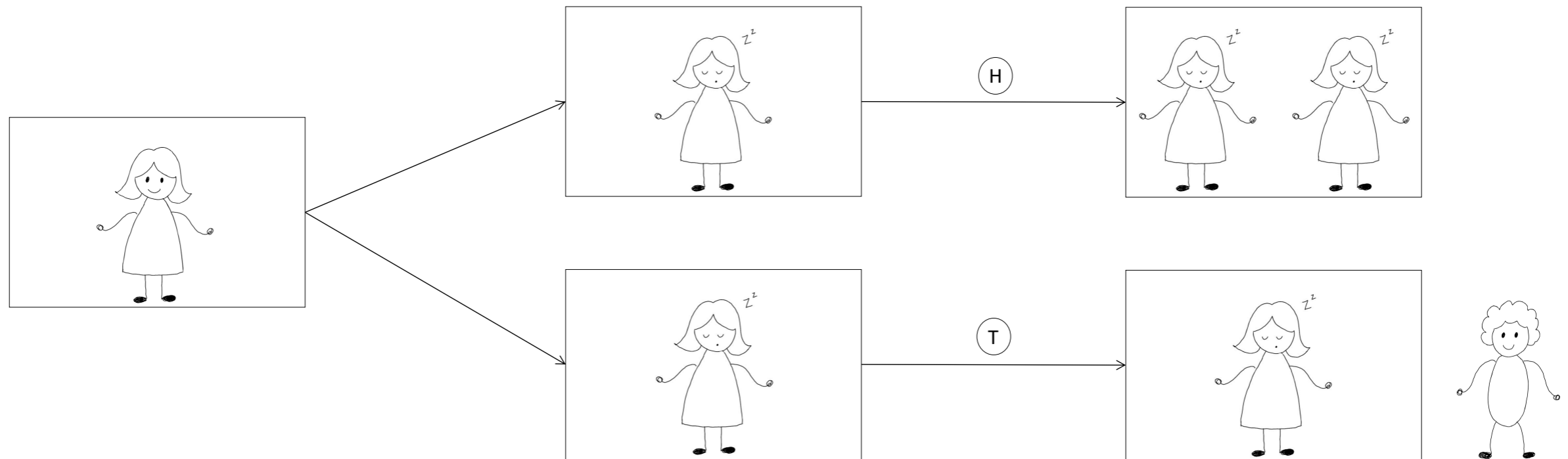
This would imply that Freya will probably continue seeing “Earth-like business as usual” even if Earth-Freya was hugely outnumbered by BBs.

To make sense of this statement, the agent “Freya” would not be defined by a localized physical system on which she supervenes, but by her **structural properties**, represented both on Earth and in the BBs.

→ hints more at a structuralist / idealist / fragmentalist view.

Overview

1. Reproducing WF phenomenology with classical **duplication** (“thinking twice inside the box”)



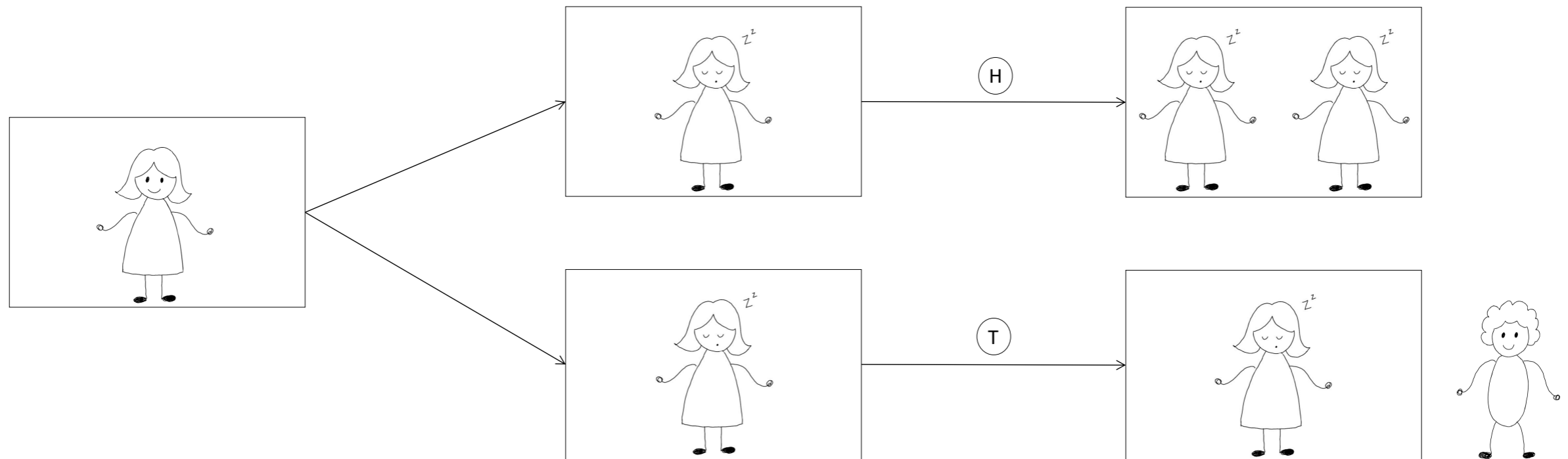
2. A common *structural* core: **Restriction A**

3. Restriction A elsewhere: **Boltzmann brain problem++**

4. Consequence: **Fragmentalism/idealism**

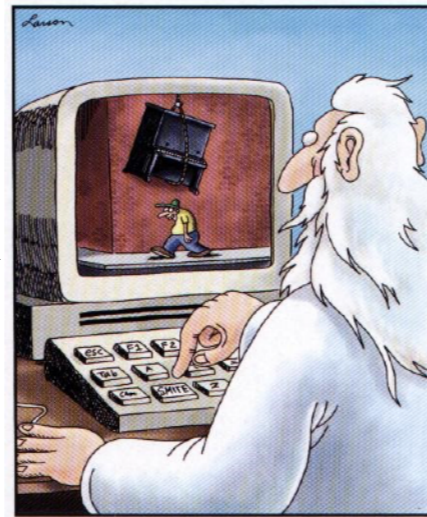
Overview

1. Reproducing WF phenomenology with classical **duplication** (“thinking twice inside the box”)

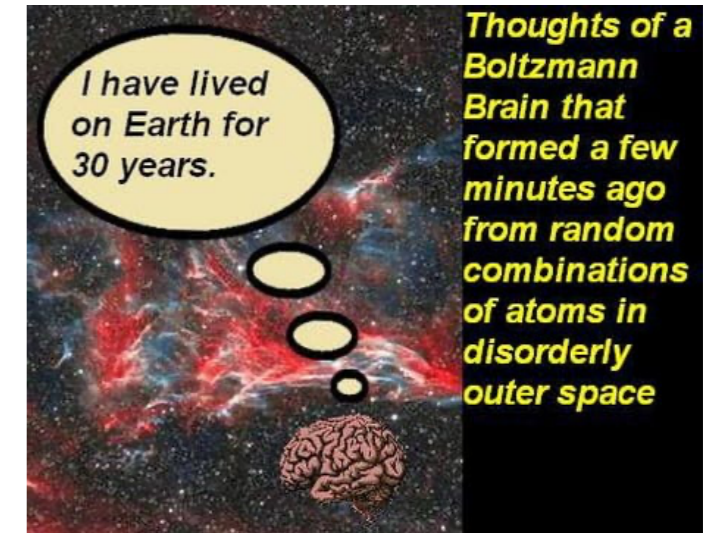


2. A common *structural* core: **Restriction A**
3. Restriction A elsewhere: **Boltzmann brain problem++**
4. Consequence: **Fragmentalism/idealism**

What about our to-be-simulated friend and the cosmologists?

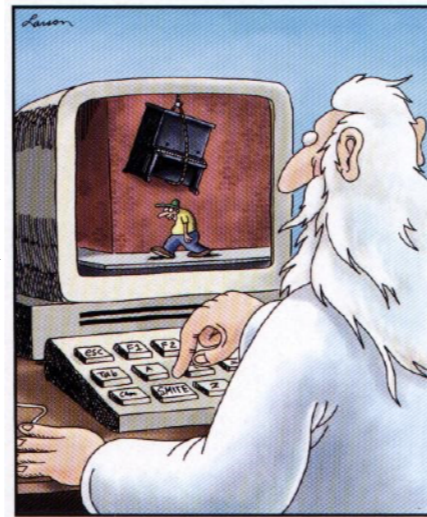


God at His computer

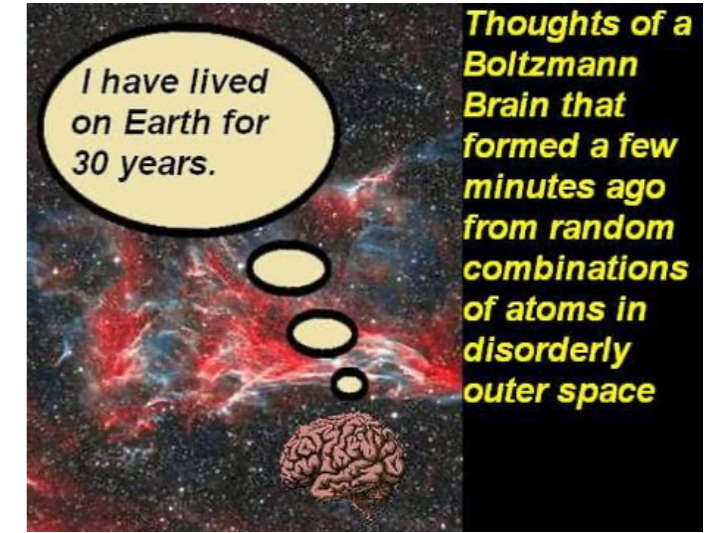


For every **single** agent, there should be a mathematical expression for what they should believe about future experiences, *in all circumstances*. In ordinary physics, this is exactly the quantum state, as we have learned yesterday from Philipp Berghofer. Should not assume, but **derive** this!

What about our to-be-simulated friend and the cosmologists?



God at His computer



For every **single** agent, there should be a mathematical expression for what they should believe about future experiences, *in all circumstances*. In ordinary physics, this is exactly the quantum state, as we have learned yesterday from Philipp Berghofer. Should not assume, but **derive** this!

- This would be a theory that explains, starting with private probabilities,
- how a notion of “external world” emerges for $N=1$ agents, and
 - how an approximate notion of objectivity emerges for $N>1$ agents.
 - Since the **personalist** probabilities do not typically fit together into a joint distribution, the hope is that aspects of quantum theory arise.
 - Then we can tell our to-be-simulated friend & ... what to expect.

Algorithmic idealism

An approach of this sort already exists, and is under further construction.

If you are interested: mpmueller.net/ai

Soon on the arXiv: Adversarial collaboration with Kelvin McQueen.

Paper “Law without law: ...”, Quantum **4**, 301 (2020).

Conclusions

- Have shown how to reproduce certain structural aspects of Wigner's friend scenarios classically via **duplication**.
- **Restriction A** as a common core: physical theories do not always give us joint probability distributions for the future observations of all agents (or even a *single* agent).
Have argued that this is what Wigner's friend is ultimately about.
- This is at the core of several other puzzles in physics and philosophy, including the **Boltzmann brain problem**.

Caroline Jones and MM, arXiv:2402.08727 (and mpmueller.net/ai)

- Have argued that this motivates idealist/fragmentalist approaches where “reality” is a mosaic of the fundamental first-person pieces.

Thank you!