

Multi-agent paradoxes beyond quantum theory

V. Vilasini¹, Nuriya Nurgalieva², and Lída del Rio²

¹Department of Mathematics, University of York, Heslington, York, YO10 5DD, UK

²Institute for Theoretical Physics, ETH Zurich, 8093 Zürich, Switzerland

Which theories lead to a contradiction between simple reasoning principles and modelling observers' memories as physical systems? Frauchiger and Renner have shown that this is the case for quantum theory [1]. Here we generalize the conditions of the Frauchiger-Renner result so that they can be applied to arbitrary physical theories, and in particular to those expressed as *generalized probabilistic theories* (GPTs) [2, 3]. We then apply them to a particular GPT, box world, and find a deterministic contradiction in the case where agents may share a PR box [4], which is stronger than the quantum paradox, in that it does not rely on post-selection. Obtaining an inconsistency for the framework of GPTs broadens the landscape of theories which are affected by the application of classical rules of reasoning to physical agents. In addition, we model how observers' memories may evolve in box world, in a way consistent with Barrett's criteria for allowed operations [3].

1 Motivation

In order to process information and make logical inferences, we would like to be able to apply simple reasoning principles to all situations. By this we mean that ideally we would like inferences such as “if I know that a holds, and I know that a implies b , then I know that b holds” to be valid independently of the nature of a and b — to take logic as a primitive that can be applied to any physical setting. When considering scenarios with several rational agents, this extends to reasoning about each other's knowledge. Examples include games like poker, complex auctions, cryptographic scenarios, and of course [logical hat puzzles](#), where we must process complex statements of the sort “I know that she knows that he does not know a ” to keep track of the flows of knowledge. On the other hand, when we describe the world through physics, we would like to consider ourselves a part of it, and in particular we would like to model our brains and memories as physical systems described by some theory. When that theory is quantum mechanics, it turns out that these two desiderata (applying to reason about each other's knowledge, and modelling memories as physical systems) are incompatible. This was first pointed out by Frauchiger and Renner, in a thought experiment where agents who can measure each others memories (modelled as quantum systems) and reason about shared and individual knowledge may reach contradictory conclusions [1]. We will not review the original experiment here; a pedagogical exposition can be found in our paper [5], but is not necessary to follow this article.

Our ultimate goal is to understand whether this incompatibility between multi-agent logic and physics is a peculiar feature of quantum theory, or if other physical theories also admit this kind of contradictions. If the latter is true, we would like to outline a class of theories where

V. Vilasini: vv577@york.ac.uk

Nuriya Nurgalieva: nuriya@phys.ethz.ch

Lída del Rio: lidia@phys.ethz.ch

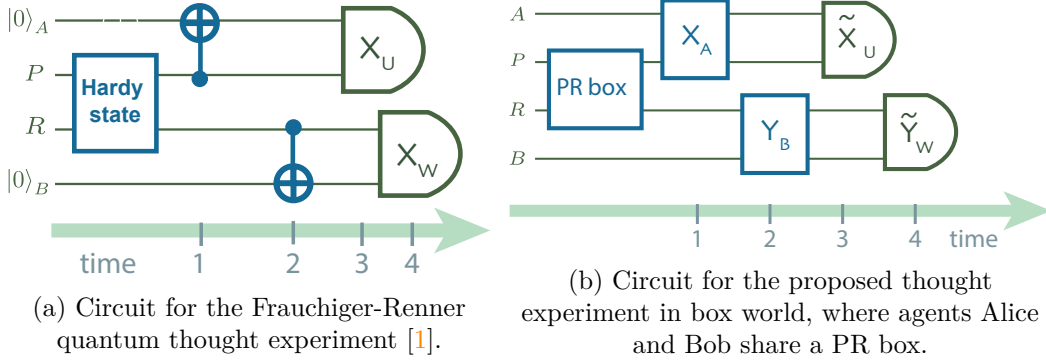


Figure 1: Circuits for multi-agent paradoxical experiments as seen by outside observers. Two instances of thought experiments which lead to a logical contradiction: in quantum theory and in box world. **a)** In the quantum case, the original experiment is formulated as a prepare and measure scenario [1] but is equivalent to the version shown here. Alice and Bob share a Hardy state, $|\Psi\rangle_{PR} = 1/\sqrt{3}(|00\rangle + |10\rangle + |11\rangle)$ where the P and R systems correspond to Alice’s and Bob’s halves of the state respectively. They then measure their halves of the state in the Z basis and record the outcome in their quantum memories A and B . For outside observers Ursula and Wigner, these measurements are modelled as unitary evolutions (like the CNOT gates pictured) that correlate memory and the subsystem measured. Finally, Ursula (and Wigner) measure AP (and RB) in the basis $\{|ok\rangle_{AP} = 1/\sqrt{2}(|00\rangle_{AP} - |11\rangle_{AP}), |fail\rangle_{AP} = 1/\sqrt{2}(|00\rangle_{AP} + |11\rangle_{AP})\}$ (and analogously defined $\{|ok\rangle_{RB}, |fail\rangle_{RB}\}$). After all is done, agents reason about each others’ knowledge to find a contradiction. In the quantum case, this only happens if Ursula and Wigner obtain the outcomes $u = w = ok$ (giving the paradoxical chain $u = w = ok \Rightarrow b = 1 \Rightarrow a = 1 \Rightarrow w = fail$ where a and b are Alice’s and Bob’s outcomes). **b)** In box world, Alice and Bob share a PR box and measure their halves using some measurements labelled $X = Y$ respectively (not to be confused with X, Y basis measurements), updating their memories A and B . Ursula and Wigner can now measure the joint systems AP and RB using different settings ($\tilde{X} = \tilde{Y} = 1 = X \oplus 1$). In this case, the contradiction can always be found independently of their outcomes as we show in Section 4.

these logical inconsistencies may arise. Such an analysis could help us identify the features of quantum theory responsible for such a paradox; in particular, here we investigate the landscape of generalized probabilistic theories [2, 3].

Contributions of this work. In Section 2, we generalize conditions on reasoning, memories and measurements so that they can be applied to any physical theory. The conditions can be briefly summarized as: agents may use logic to reason about each others’ knowledge; a physical theory allows agents to make predictions about the outcomes of measurements; and a measurement by an agent Alice may be modelled by others as a physical evolution on her lab which preserve the information about the original system measured (from the outside agents’ perspective). This generalizes the von Neumann view of measurements as a unitary evolution of the system and measurement apparatus [6]. In Section 3 we apply those conditions to the framework of *generalized probabilistic theories* (GPTs) [2, 3]; in particular we introduce a way to describe an agent’s measurement from the perspective of other agents. Finally, in Section 4 we derive a logical inconsistency akin to one found in [1], using a setup where agents share a PR box, a maximally non-local resource in box-world (a GPT). The paradox found is stronger than the quantum one, in the sense that it does not rely on post-selection: agents always reach a

contradiction, independently of the outcome¹. A high-level circuit representation of the original experiment, as well as the PR box version, are depicted in Figure 1.

2 Generalized reasoning, memories and measurements

Here we generalize the Frauchiger-Renner conditions for inter-agent consistency to general physical theories. The conditions can be instantiated by each specific theory. This includes but is not limited to theories framed in the approach of generalized probabilistic theories [2]. In some theories, like quantum mechanics and box world (a GPT), we will find these four conditions to be incompatible, by finding a direct contradiction in examples like the Frauchiger-Renner experiment or the PR-box experiment described in Section 3. In other theories (like classical mechanics and Spekkens’ toy theory [9]) these four conditions may be compatible. A complete characterization of theories where one can find these paradoxes is the subject of future work.

2.1 Reasoning about knowledge

This condition is theory-independent. It tells us that rational agents can reason about each others knowledge in the usual way. This is formalized by a weaker version of *epistemic modal logic*, which we explain in the following (for the full derivation of the form used here see [5]).

Let us start with a simple example. The goal of modal logic is to allow us to operate with chained statements like “Alice knows that Bob knows that Eve doesn’t know the secret key k , and Alice further knows that $k = 1$,” which can be expressed as

$$K_A [(K_B \neg K_E k) \wedge k = 1],$$

where the operators K_i stand for “agent i knows.” If in addition Alice trusts Bob to be a rational, reliable agent, she can deduce from the statement “I know that Bob knows that Eve doesn’t know the key” that “I know that Eve doesn’t know the key”, and forget about the source of information (Bob). This is expressed as

$$K_A(K_B \neg K_E k) \implies K_A \neg K_E k.$$

We should also allow Alice to make deductions of the type “since Eve does not know the secret key, and one would need to know the key in order to recover the encrypted message m , I conclude that Eve cannot know the secret message,” which can be encoded as

$$K_A[(\neg K_E k) \wedge (K_i m \implies K_i k, \forall i)] \implies K_A \neg K_E m.$$

Generalizing from this example, this gives us the following structure.

Definition 1 (Reasoning agents) *An experimental setup with multiple agents A_1, \dots, A_N can be described by knowledge operators K_1, \dots, K_N and statements $\phi \in \Phi$, such that $K_i \phi$ denotes “agent A_i knows ϕ .” It should allow agents to make deductions (Figure 2a), that is*

$$K_i[\phi \wedge (\phi \implies \psi)] \implies K_i \psi.$$

Furthermore, each experimental setup defines a trust relation between agents (Figure 2b): we say that an agent A_i trusts another agent A_j (and denote it by $A_i \rightsquigarrow A_j$) iff for all statements ϕ , we have

$$K_i(K_j \phi) \implies K_i \phi.$$

¹The joint state and the probability distributions of the original Frauchiger-Renner paradox are akin to those of Hardy’s paradox [7]. For a comparison of Hardy’s paradox and PR box and why the latter allows for a contradiction without post-selection, see [8].



(a) An agent using deduction, applying a modal logic distribution axiom (for a short review of modal logic framework and axioms, check Appendix A).

(b) An agent A_i trusts another agent A_j , denoted by $A_j \rightsquigarrow A_i$.

Figure 2: **An illustrated example of reasoning agents of Definition 1.** Here we can see that agents are allowed to make deductions and transfer knowledge from one another, given the respectful trust relation.

For the purposes of following the example of Section 4, this informal definition suffices. The full formal version of the axioms of modal logic used here can be found in Appendix A.²

2.2 Physical theories and agents as physical systems

This condition is to be instantiated by each physical theory, and is the way that we incorporate the physical theory into the reasoning framework used by agents in a given setting. If all agents use the same theory to model the operational experiment (like quantum mechanics, special relativity, classical statistical physics, or box world), this is included in the *common knowledge* shared by the agents. For example, in the case of quantum theory, we have that “everyone knows that the probability of obtaining outcome $|x\rangle$ when measuring a state $|\psi\rangle$ is given by $|\langle x|\psi\rangle|^2$, and everyone knows that everyone knows this, and so on.”

Definition 2 (Common knowledge) *We model a physical theory shared by all agents $\{A_i\}_i$ in a given setting as a set \mathbb{T} of statements that are common knowledge shared by all agents, i.e.*

$$\phi \in \mathbb{T} \iff (\{K_i\}_i)^n \phi, \quad \forall n \in \mathbb{N},$$

where $(\{K_i\}_i)^n$ is the set of all possible sequences of n operators picked from $\{K_i\}_i$. For example, $(K_1 K_5 K_1 K_2) \in (\{K_i\}_i)^4$ and stands for “agent A_1 knows that agent A_5 knows that agent A_1 knows that agent A_2 knows.”

²Note that in general ‘one human \neq one agent.’ For example, consider a setting where we know that Alice’s memory will be tampered with at time τ (much like the original Frauchiger-Renner experiment, or the sleeping beauty paradox [10]). We can define two different agents $A_{t<\tau}$ and $A_{t>\tau}$ to represent Alice before and after the tampering — and then for example Bob could trust pre-tampering (but not post-tampering) Alice, $A_{t<\tau} \rightsquigarrow B$.



Figure 3: **Common knowledge.** An example of what common knowledge (in this case, a shared physical theory \mathbb{T}) is: all agents know that all agents know that ... (and so on) ... that theory \mathbb{T} holds.

Note that the set \mathbb{T} of common knowledge may include statements about the settings of the experiment, as well as complex derivations³. To find our paradoxical contradiction, we may only need a very weak version of a full physical theory: for example Frauchiger and Renner only require a possibilistic version of the Born rule, which tells us whether an outcome will be observed with certainty [1]. This will also be the case in box world.

In operational experiments, a reasoning agent can make statements about systems that she studies; consequently, the theory used by the agent must be able to produce a description or a model of such a system, namely, in terms of a set of states. For example, in quantum theory a two-state quantum system with a ground state $|0\rangle$ and an excited state $|1\rangle$ (*qubit*) can be fully described by a set of states $\{|\psi\rangle\}$ in a Hilbert space \mathcal{H} , where $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$ with $\alpha, \beta \in \mathbb{C}$ and $|\alpha|^2 + |\beta|^2 = 1$. Another examples of theories and respective descriptions of states of systems include: GPTs, where e.g. a generalised bit (*gbit*) is a system completely characterized by two binary measurements which can be performed on it [3] (a review of GPTs can be found in Section 3); algebraic quantum mechanics, with states defined as linear functionals $\rho : A \rightarrow \mathbb{C}$, where A is a C^* -algebra [6]; or resource theories with some state space Ω [11].

Definition 3 (Systems) Here we call a “physical system” (or simply “system”) anything that can be an object of a physical study⁴. A system can be characterized, according to the theory \mathbb{T} , by a set of states $\{P_S^i\}_{i \in \mathcal{I}_S}$ ($\mathcal{I}_S \subseteq \mathbb{N}$).

We have already used knowledge operators K_i to denote knowledge of each agent. Now let us add memory to the formal description of an agent.

³One can also alternatively model a physical theory as a subset \mathbb{T}_P of the set \mathbb{T} of common knowledge, $\mathbb{T}_P \subseteq \mathbb{T}$, in the case when details of experimental setup are not relevant to the theoretical formalism.

⁴We strive to be as general as possible and do not suppose or impose any structure on systems and connections between them; in particular, we don’t make any assumptions about how composite systems are formally described in terms of their parts.

Definition 4 (Agents) A physical setting may be associated with a set \mathcal{A} of agents. An agent $A_i \in \mathcal{A}$ is described by a knowledge operator $K_i \in \mathcal{K}_{\mathcal{A}}$ and a physical system $M_i \in \mathcal{M}_{\mathcal{A}}$, which we call a “memory.” Each agent may study other systems according to the theory \mathbb{T} . An agent’s memory M_i records the results and the consequences of the studies conducted by A_i . The memory may be itself an object of a study by other agents.

2.3 Measurements and memory update

Here we consider measurements both from the perspective of an agent who performs them, as well as from the point of view of another agent who is modeling the first agent’s memory.

In an experiment involving measurements, each agent has the subjective experience of only observing one outcome (independently of how others may model her memory), and we can see this as the definition of a measurement: if there is no subjective experience of observing a single outcome, we don’t call it a measurement. We can express this experience as statements such as $\phi_0 =$ “The outcome was 0, and the system is now in state $|0\rangle$.” Let us explain further after the formal definition.

Definition 5 (Measurements) A measurement is a type of study that can be conducted by an agent A_i , while studying a system S ; the essential result of the study is the obtained “outcome” $x \in \mathcal{X}_S$. If witnessed by another agent A_j (who knows that A_i performed the measurement but does not know the outcome), the measurement is characterized by a set of propositions $\{\phi_x\} \in \Phi$, where ϕ_x corresponds to the outcome x , satisfying:

- $K_j(\exists x \in \mathcal{X}_S : K_i \phi_x)$,
- $K_j K_i \phi_x \implies K_j K_i \neg(\phi_y), \quad \forall y \neq x.$

The first condition tells us that A_j knows that A_i must have observed one outcome, and derived all the relevant conclusions, as expressed by one of the propositions ϕ_x . For example, if the measurement represents a perfect Z measurement of a qubit, ϕ_0 may include statements like “the qubit is now in state $|0\rangle$; before the measurement it was not in state $|1\rangle$; if I measure it again in the same way, I will obtain outcome 0;” and so on. The second condition roughly implements the experience of observing a single outcome and trusting that information. If A_i observes x , they conclude that the conclusions ϕ_y that they would have derived had they observed a different outcome y are not valid. In the previous example, they would know that it does not hold $\phi_1 =$ “the qubit is now in state $|1\rangle$; before the measurement it was not in state $|0\rangle$; if I measure it again I will see outcome 1.” This condition also ensures that the conclusions $\{\phi_x\}_x$ are mutually incompatible, i.e. that the measurement is tightly characterized.

A measurement of another agent’s memory is also an example of a valid measurement. In other words, agent A_j can choose A_i ’s lab, consisting of A_i ’s memory and another system S (which A_i studies), as an object of her study.

Thus, any agent’s memory can be modelled by the other agents as a physical system undergoing an evolution that correlates it with the measured system. In quantum theory, this corresponds to the unitary evolution

$$\left(\sum_{x=0}^{N-1} p_x |x\rangle_{\text{system}} \right) \otimes |0\rangle_{\text{memory}} \rightarrow \sum_{x=0}^{N-1} p_x \underbrace{|x\rangle_{\text{system}} \otimes |x\rangle_{\text{memory}}}_{=: |\bar{x}\rangle_{SM}}. \quad (1)$$

The key aspect here is that the set of states of the joint system of observed system and memory, $\{P_{SM}^l\}_l = \text{span}\{|x\rangle_{\text{system}} \otimes |x\rangle_{\text{memory}}\}_{x=0}^{N-1}$ is post-measurement isomorphic to the the set of

states $\{P_S^j\}_j$ system alone. That is, for every transformation ϵ_S that you could apply to the system before the measurement, there is a corresponding transformation ϵ_{SM} acting on the $\{P_{SM}^l\}_l$ that is operationally identical. By this we mean that an outside observer would not be able to tell if they are operating with ϵ_S on a single system before the measurement, or with ϵ_{SM} on system and memory after the measurement. In particular, if ϵ_S is itself another measurement on S within a probabilistic theory, it should yield the same statistics as post-measurement ϵ_{SM} . For a quantum example that helps clarify these notions, consider S to be a qubit initially in an arbitrary state $\alpha|0\rangle_S + \beta|1\rangle_S$. An agent Alice measures S in the Z basis and stores the outcome in her memory A . While she has a subjective experience of seeing only one possible outcome, an outside observer Bob could model the joint evolution of S and A as

$$(\alpha|0\rangle_S + \beta|1\rangle_S) \otimes |0\rangle_A \rightarrow \alpha|0\rangle_S|0\rangle_A + \beta|1\rangle_S|1\rangle_A.$$

Suppose now that (before Alice’s measurement) Bob was interested in performing an X measurement on S . This would have been a measurement with projectors $\{|+\rangle\langle+|_S, |-\rangle\langle-|_S\}$, where $|\pm\rangle_S = \frac{1}{\sqrt{2}}(|0\rangle_S \pm |1\rangle_S)$. However, he arrived too late: Alice has already performed her Z measurement on S . If now Bob simply measured X on S he would obtain uniform statistics, which would be uncorrelated with the initial state of S . So what can he do? It may not be very friendly, but he can measure S and Alice’s memory A jointly, by projecting onto

$$\begin{aligned} |+\rangle_{SA} &= \frac{1}{\sqrt{2}}(|0\rangle_S|0\rangle_A + |1\rangle_S|1\rangle_A) \\ |-\rangle_{SA} &= \frac{1}{\sqrt{2}}(|0\rangle_S|0\rangle_A - |1\rangle_S|1\rangle_A), \end{aligned}$$

which yields the same statistics of Bob’s originally planned measurement on S , had Alice not measured it first. This equivalence should also hold in the more general case where the observed system may have been previously correlated with some other reference system: such correlations should be preserved in the measurement process, as modelled from the “outside” observer Bob.

There are many options to formalize this notion that “every way that an outside observer could have manipulated the system before the measurement, he may now manipulate a subspace of ‘system and observer’s memory,’ with the same results.” A possible simplification to restrict our options is to take subsystems and the tensor product structure as primitives of the theory, which is the case for GPTs [3] but not for general physical theories (like field theories; for a discussion see [12]). In the interest of time, we will for now restrict ourselves to this case, and leave a more general formulation of this condition as future work. For simplicity, we also restrict ourselves to information-preserving measurements (excluding for now those where some information may have leaked to an environment external to Alice’s memory), which are sufficient to derive the contradiction.

Definition 6 (Information-preserving memory update) *Let $\{P_S^j\}_j$ be a set of states of a system S that is being studied by an agent A_i with a memory M_i , and $\{P_{SM_i}^l\}_l$ be a set of states of the joint system SM_i , which consists of the systems S and M_i . Then a map $u : \{P_S^j\}_j \rightarrow \{P_{SM_i}^l\}_l$ is called an information-preserving memory update if for all operations $\mathcal{E}_S : \{P_S^j\}_j \rightarrow \{P_S^j\}_j$ on the system S , there exists an operation $\mathcal{E}_{SM_i} : \{P_{SM_i}^l\}_l \rightarrow \{P_{SM_i}^l\}_l$ such that:*

$$\forall P_S^i, \forall A_j \in \mathcal{A} \quad K_j \phi[\mathcal{E}_S(P_S^i)] \Rightarrow K_j \phi[\mathcal{E}_{SM_i} \circ u(P_S^i)].$$

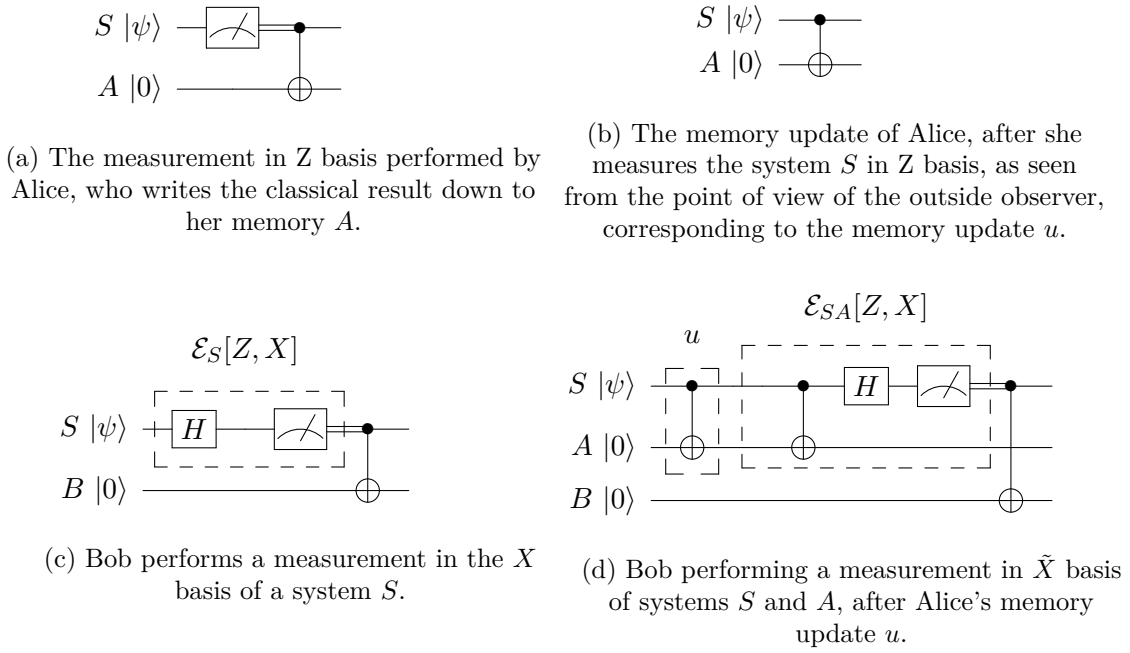


Figure 4: The measurement and memory update in quantum theory from different perspectives. From Alice's point of view, the measurement of the system S either in Z basis yields a classical result, which she records to her memory A , performing a classical CNOT (Figure 4a). From an outside observer, Bob's perspective, as he is not aware of Alice's measurement result, the CNOT is a quantum entangling operation, which corresponds to the memory update u (Figure 4b). If he had access to the system S prior to the measurement by A , and wanted to measure it in X basis ($\{|+\rangle_S, |-\rangle_S\}$), he would have to perform an operation $\mathcal{E}_S[Z, X]$ (and then copy the classical result into his memory B) (Figure 4c). If the system S was initially in a state $|\psi\rangle = |+\rangle_S$, then a proposition which would correspond to this operation is $\phi[\mathcal{E}_S[Z, X](|\psi\rangle_S)] = "s = +"$. However, if the measurement in Z is already performed by A and the result is written to her memory, the whole process described by Bob as a memory update u , and in order to comply his initial wish to measure S only, he can perform an operation $\mathcal{E}_{SA}[Z, X]$ on S and A together instead, which is a measurement in $\{|+\rangle_{SA}, |-\rangle_{SA}\}$ basis (Figure 4d). A proposition which this operation yields is $\phi[\mathcal{E}_{SM_i} \circ u(|\phi\rangle_S)] = "sa = +"$ (as $|\xi\rangle_{SA} = |+\rangle_{SA}$), which naturally follows from " $s = +$ ", given the structure of the memory update u .

See Figure 4 for an example. In general, the memory update map u need not be reversible; for example, in box world it is an irreversible transformation, as we will see later.

Note that the characterization of measurements introduced in this section is rather minimal. In physical theories like classical and quantum mechanics, measurements have other natural properties that we do not require here. Two striking examples are "after her measurement, Alice's memory becomes correlated with the system measured in such a way that, for any subsequent operation that Bob could perform on the system, there is an equivalent operation he may perform on her memory" and "the correlations are such that there exists a joint operation on the system and Alice's memory that would allow Bob to conclude which measurement Alice performed." While these properties hold in the familiar classical and quantum worlds, we do not know of other physical theories where measurements can satisfy them, and they require Bob to be able to act independently on the system and on Alice's memory, which may not always be possible. For example, we will see that in box world, these two subsystems become *superglued*

after Alice’s measurement, and that Bob only has access to them as a whole and not as individual components. As such, we will not require these properties out of measurements, for now. We revisit this discussion in Section 5.

3 Box world: states and memories

Generalised probabilistic theories [2, 3] (GPTs) provide an operational framework for describing probabilistic theories, including classical and quantum theories where the physical systems are taken as black boxes, characterized only by their input and output behaviour. The *state* of a system is represented by a probability vector \mathbf{P} that encodes the probabilities of possible outcomes given all the possible choices of measurement. This is a single-shot characterization of a system: the post-measurement state can be represented by a new probability vector, and the update rules depend on the specific theory.

In this paper, we employ the framework for information processing in GPTs presented by Barrett in [3], and use the term “box world” to denote the set of theories that Barrett originally calls *Generalised No-Signalling Theories*. We will derive the paradox in box world, which is a particular instance of a GPT. However, the general assumptions proposed in Section 2 can also be applied to more general GPTs that do not obey the standard no signalling principle [13, 14] or that which obey different physical principles. We present here the minimal formalism needed to follow the argument; see Appendix B for more details.

3.1 States and operations (review)

Individual states. The so-called generalised bit or *gbit* is a system completely characterized by two binary measurements which can be performed on it [3]. Such sets of measurements that completely characterise the state of a system are known as *fiducial measurements*. The state of a gbit is thus fully specified by the vector

$$\mathbf{P}_{gbit} = \begin{pmatrix} P(a = 0|X = 0) \\ P(a = 1|X = 0) \\ P(a = 0|X = 1) \\ P(a = 1|X = 1) \end{pmatrix}, \quad (2)$$

where $X = 0$ and $X = 1$ represent the two choices of measurements and $a \in \{0, 1\}$ are the possible outcomes (Figure 5a). Analogously, a classical bit is a system characterized by a single binary fiducial measurement,

$$\mathbf{P}_{bit} = \begin{pmatrix} P(a = 0|X = 0) \\ P(a = 1|X = 0) \end{pmatrix}, \quad (3)$$

and, in quantum theory, a qubit is characterized by three fiducial measurements (corresponding, for example, to three directions X , Y and Z in the Bloch sphere),

$$\mathbf{P}_{qubit} = \begin{pmatrix} P(a = 0|X = 0) \\ P(a = 1|X = 0) \\ P(a = 0|X = 1) \\ P(a = 1|X = 1) \\ P(a = 0|X = 2) \\ P(a = 1|X = 2) \end{pmatrix}. \quad (4)$$

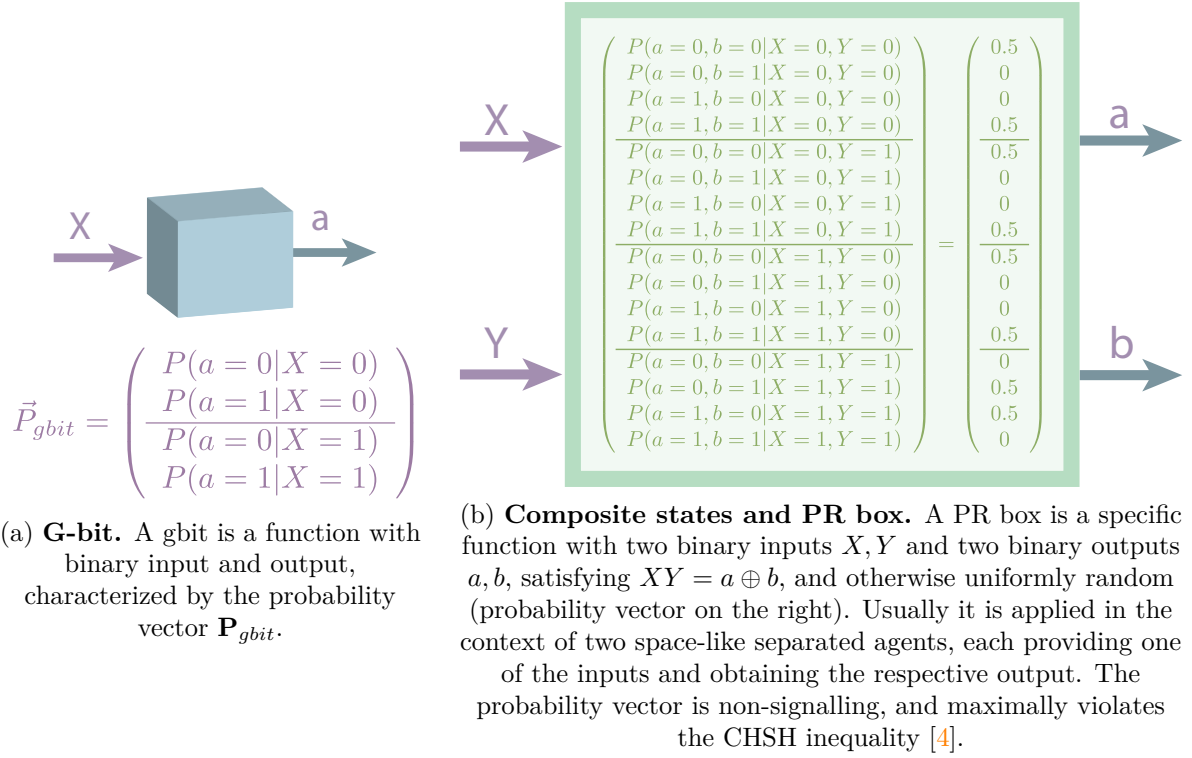


Figure 5: **Boxes in Generalized Probabilistic Theories.** The modular objects of GPTs are input/output functions depicted as boxes and characterized by probability vectors. Each function (or box) can be evaluated once, and it may or not correspond to a physical system being probed; even if it is, nothing is assumed about the post-evaluation state of the system (unlike quantum theory, which specifies the post-measurement state of a system given its initial state and the measurement device).

For normalized states, we have $|\mathbf{P}| = \sum_i P(a = i|X = j) = 1, \forall j$. The set of possible states of a gbit is convex, with extremes

$$\mathbf{P}_{00} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{P}_{01} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{P}_{10} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{P}_{11} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}. \quad (5)$$

These correspond to pure states. In the qubit case, the extremes correspond to all the points on the surface of the Bloch sphere, for example

$$\mathbf{P}_{|+\rangle} = \begin{pmatrix} 1 \\ 0 \\ 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{pmatrix}, \quad \mathbf{P}_{|-\rangle} = \begin{pmatrix} 0 \\ 1 \\ 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{pmatrix}, \quad \mathbf{P}_{|0\rangle} = \begin{pmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \\ 1 \\ 0 \end{pmatrix}, \quad \mathbf{P}_{|1\rangle} = \begin{pmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \\ 0 \\ 1 \end{pmatrix}. \quad (6)$$

Note that in box world, pure gbits are deterministic for both alternative measurements, whereas in quantum theory at most one fiducial measurement can be deterministic for each pure qubit, as reflected by uncertainty relations. We denote the set of allowed states of a system A by \mathcal{S}^A .

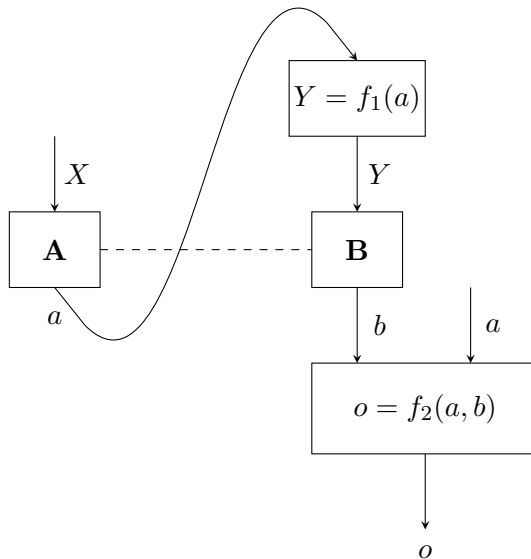


Figure 6: **Bipartite measurements in boxworld.** Any bipartite measurement on a 2-gbit box world system can be decomposed into a procedure (or convex combinations thereof) of the following form. Alice first performs a measurement X on one of the gbits (labelled A), and forwards the outcome a to Bob. Bob then performs a measurement $Y = f_1(a)$, which may depend on a , on the other gbit (labelled B), obtaining the outcome b . The final measurement outcome o of the joint measurement can be computed by Bob as a function $f_2(a, b)$. All allowed bipartite measurements are convex combinations of this type of classical wirings [3].

Composite states. The state of a bipartite system AB , denoted by $\mathbf{P}^{AB} \in \mathcal{S}^{AB}$ can be written in the form $\mathbf{P}^{AB} = \sum_i r_i \mathbf{P}_i^A \otimes \mathbf{P}_i^B$ where r_i are real coefficients⁵ and $\mathbf{P}_i^A \in \mathcal{S}^A$, $\mathbf{P}_i^B \in \mathcal{S}^B$ can be taken to be pure and normalised states of the individual systems A and B [3]. Thus, a general 2-gbit state \mathbf{P}_2^{AB} can be written as in Figure 5b (left), where $X, Y \in \{0, 1\}$ are the two fiducial measurements on the first and second gbit and $a, b \in \{0, 1\}$ are the corresponding measurement outcomes. The PR box \mathbf{P}_{PR} , on the right, is an example of such a 2 gbit state that is valid in box world, which satisfies the condition $a \oplus b = xy$ [4].

State transformations. Valid operations are represented as matrices that transform valid state vectors to valid state vectors (Appendix B). In addition, we only have access to the (single-shot) input/output behaviour of systems, so in practice all valid operations in box world take the form of classical wirings between boxes, which correspond to pre- and post-processing of input and output values, and convex combinations thereof [3]. For example, bipartite joint measurements on a 2-gbit system can be decomposed into convex combinations of classical “wirings”, as shown in Figure 6. In contrast, quantum theory allows for a richer structure of bipartite measurements by allowing for entangling measurements (e.g. in the Bell basis), which cannot be decomposed into classical wirings. Bipartite transformations on multi-gbit systems turn out to be classical wirings as well [3]. Reversible operations in particular consist only of trivial wirings: local operations and permutations of systems [15]. One cannot perform entangling operations such as a coherent copy (the quantum CNOT gate) [3, 16], which is

⁵Note that it is not necessary that the coefficients r_i be positive and sum to one. If this is the case, then the composite state would be separable and hence local, otherwise, the state is entangled [3].

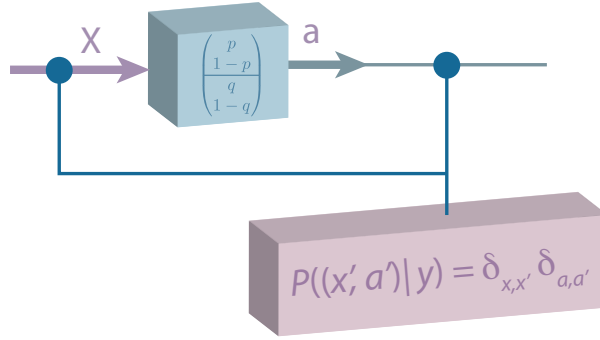


Figure 7: **Measurement: observer’s perspective.** An agent Alice measures a system with measurement setting X , and obtains outcome a with a given probability. In the language of GPTs this corresponds to running the box that encodes the measurement statistics. Alice may then the measurement data (input and output) to memory. If this is a classical memory, like a notebook, the procedure corresponds to preparing a new box (to be run later by herself), which outputs the pair (X, a) deterministically.

required in the original version of the Frauchiger-Renner experiment.

3.2 Agents, memory and measurement in box world

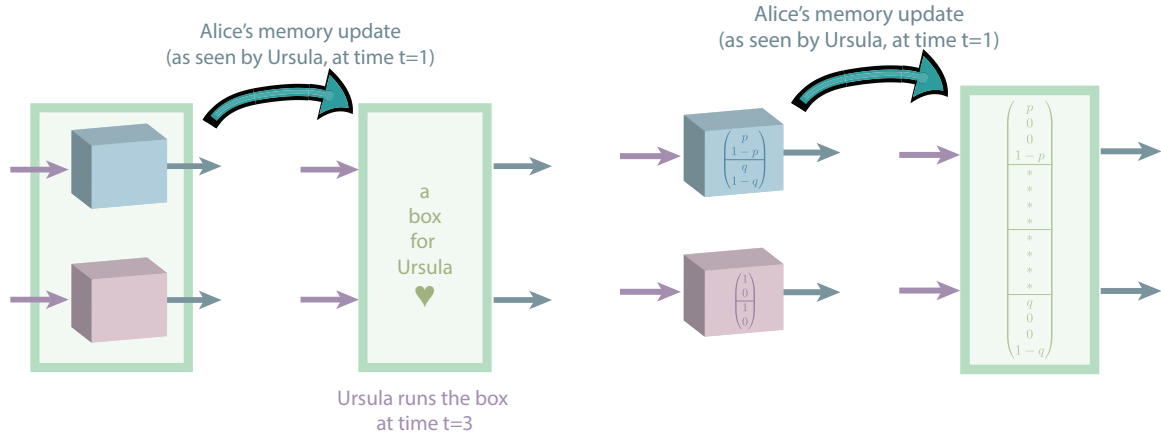
We will now instantiate our general conditions for agents, memories and measurements (definitions definitions 4 to 6) in box world. As there is no physical theory for the dynamics behind box world, there is plenty of freedom in the choice of implementation. In principle each such choice could represent a different physical theory leading to the same black-box behaviour in the limit of a single agent with an implicit memory. This is analogous to the way in which different versions of quantum theory (Bohmian mechanics, collapse theories, unitary quantum mechanics with von Neumann measurements) result in the same effective theory in that limit.

Definition 7 (Agents in box world) *Let \mathbb{T} be the theory that describes box world, according to [3]. As per definition 4, an agent $A_i \in \mathcal{A}$ is described by a knowledge operator $K_i \in \mathcal{K}_{\mathcal{A}}$ and a physical memory $M_i \in \mathcal{M}_{\mathcal{A}}$.*

We will focus on the case where the memory consists of bit or qbits. Each agent may study other systems according to the theory \mathbb{T} . An agent’s memory records the results and the consequences of the studies conducted by them, and may be an object of a study by other agents.

It is worth mentioning that boxes do not correspond to physical systems, but to input/output functions that can only be evaluated once. As such, the post-measurement state of a physical system is described by a whole new box. The notion of an individual system itself, as we will see, may be unstable under measurements — some measurements *glue* the system to the observer’s memory, in a way that makes individual access to the original system impossible.

Measurement: observer’s perspective. From the point of view of the observer who is measuring (say Alice), making a measurement on a system corresponds simply to running the box whose state vector encodes the measurement statistics. Alice may then commit the result of her measurement to a physical memory, like a notebook where she writes ‘I measured observable X and obtained outcome a .’ To be useful, this should be a memory that may be consulted later, i.e. it could receive an input Y = ‘open and read the memory’, and output the pair (X, a) . In the language of GPTs, this means that Alice, from her own perspective, prepares a new box with one



(a) Generally, in GPTs with some notion of subsystems, Ursula can think of the physical system measured by Alice, and Alice's memory pre-measurement as two boxes, which Ursula could in principle run if Alice chose not to measure (left). For example, the statistics of the system could correspond to its initial state, and the statistics of the memory to a standard, reset gbit. From Ursula's perspective, Alice's measurement corresponds to some transformation that results on a final state on which Ursula can later act. This final state can be represented by a new box available to Ursula, which will have in principle a different behaviour, depending on the concrete physical theory.

(b) In the specific case of box world, if the two initial systems correspond to small gbit boxes, and Alice's memory is initialized as shown, and if we want to preserve the global system dimensions, then the rules for allowed transformations limit the statistics of Ursula's final box to be of the form shown in the right (Appendix C). The asterisks represent arbitrary values, which will depend on the choice of implementation of Alice's measurement. This transformation is in principle non-reversible: note that in the final box, Ursula cannot address system and memory independently, but only the global, *superglued* box.

Figure 8: **Memory update after a measurement: an outsider's perspective.** Here Alice makes a measurement of a system (blue, top) at time $t = 1$ and stores her outcome in her memory (pink, bottom). The question is how an outsider, Ursula, models Alice's measurement. In particular, what can Ursula do with the post-measurement state?

input Y and two outputs (X', a') , with the behaviour $\mathbf{P}((X', a')|Y) = \delta_{X,X'}\delta_{a,a'}$, which depends on her observations (Figure 7). She may later run this box (look at her notebook) and recover the measurement data. The exact dimension of the box will depend on how Alice perceives and models her own memory; for example it could consist of two bits, or two gbits, or, if we think that before the measurement she stored the information about the choice of observable elsewhere, it could be a single bit or gbit encoding only the outcome. We leave this open for now, as we do not want to constrain the theory too much at this stage.

Measurements: inferences. To see the kind of inferences and conclusions that an agent can take from a measurement in box world, it's convenient to look at the example where Alice and Bob share a PR box. Suppose that Alice measured her half of the box with input $X = 1$ and obtained outcome $a = 0$. From the PR correlations, $XY = a \oplus b$, she can conclude that if Bob measures $Y = 0$, he must obtain $b = 0$, and if he measures $Y = 1$, he must obtain $b = 1$. This is independent of whether Bob's measurement happens before or after Alice (or even space-like separated). She could reach similar deterministic conclusions for her other choice

of measurement and possible outcomes. In the language of Definition 5, we have

$$\begin{aligned}\phi_{X=0,a=0} &= "[Y = 0 \implies b = 0] \wedge [Y = 1 \implies b = 0]", \\ \phi_{X=0,a=1} &= "[Y = 0 \implies b = 1] \wedge [Y = 1 \implies b = 1]", \\ \phi_{X=1,a=0} &= "[Y = 0 \implies b = 0] \wedge [Y = 1 \implies b = 1]", \\ \phi_{X=1,a=1} &= "[Y = 0 \implies b = 1] \wedge [Y = 1 \implies b = 0].\end{aligned}$$

Measurement: memory update from an outsider’s perspective. Next we need to model how an outsider agent, Ursula, models Alice’s measurement, in the case where Alice does not communicate her outcome to Ursula.⁶ Suppose that all agents share a time reference frame, and Alice makes her measurement at time $t = 1$. From Ursula’s perspective, in the most general case, this will correspond to Alice preparing a new box, with some number of inputs and outputs, which Ursula can later run (Figure 8a). The exact form of this box will depend on the underlying physical theory for measurements: in the quantum case it corresponds to a box with the measurement statistics of a state that’s entangled between the system measured and Alice’s memory, as we saw. In classical mechanics, it will correspond to perfect classical correlations between those two subsystems. In a theory of very destructive measurements, it could be that Alice’s post-measurement state is trivial from the point of view of Ursula and the resulting box is void. Now suppose that we would like to have a physical theory where the dimension of systems is preserved by measurements: for example, if the system that Alice measures is instantiated by a box with binary input and output (e.g. a gbit, or half of a PR-box), and Alice’s memory, where she stores the outcome of the measurement (as in Figure 7) is also represented as a gbit, then we would want the post-measurement box accessible to Ursula to have in total two binary inputs and two binary outputs (or more generally, four possible inputs and four possible outputs). Note that this is not a required condition for a theory to be *physical* per se — it is just a familiar rule of thumb that gives some persistent meaning to the notion of subsystems and dimensions. In such a theory that supports box world correlations, we find that the allowed statistics of Ursula’s box must satisfy the conditions of Figure 8b (proof in Appendix C). These conditions still leave us some wiggle room for possible different implementations.

Measurements: information-preserving memory update. In order to find a multi-agent paradox, we will need a model of memory update that is information-preserving, in the sense of Definition 6. This does not imply that Alice’s transformation (as seen by Ursula) be reversible: in fact, it will *glue* two subsystems such that Ursula will only be able to address them as a whole, but the relevant fact is that Ursula can apply some post-processing in order to obtain a new box with the same behaviour as the pre-measurement system that Alice observed. In Figure 9 we give an example of a model that satisfies these conditions, in addition to the conditions of Figure 8b. As we foreshadowed, this model is not completely satisfying from a physical standpoint: it looks rather trivial (a post-processing of classical outputs); the super-gluing is postulated rather than naturally emergent; and, unlike quantum von-Neumann measurements, it does not give Ursula information about the nature of Alice’s measurement. It helps to think of it as one minimal implementation among many possible, which already allows us to derive a paradox. We discuss these limitations and alternatives in Section 5.2. What is important here (and proven in Appendix C) is that this model generalizes to the case where Alice measures half of a bipartite state, like a PR box. That is, suppose that Alice and Bob share a PR box. Imagine that at time $t = 1$ Alice makes her measurement X , obtaining (from her perspective) an outcome

⁶Naming convention: as we will see in Section 4, in the proposed experiment we have two “internal” agents, Alice and Bob, who will in turn be measured by two “external” agents, Ursula and Wigner respectively. Ursula is named after Le Guin. In the example of Section 2 the internal agent was Alice and the external Bob, so that their different pronouns could help keep track of whose memory we were referring to, but we trust that the reader has got a handle on it by now.

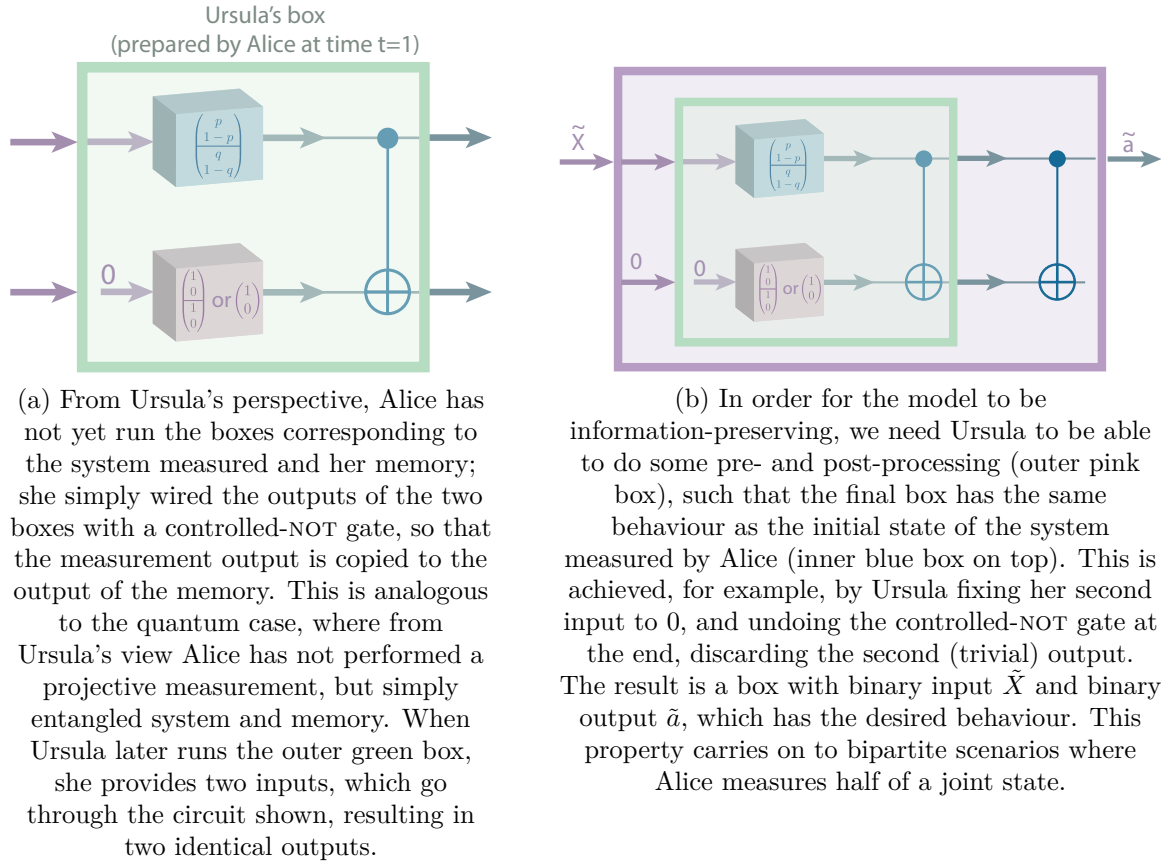


Figure 9: **Information-preserving memory update.** This (trivial) physical implementation of Alice's measurement in box world satisfies the conditions of Figure 8b and is information-preserving, in the sense that an external agent, Ursula, can run the final box as if it were the original, pre-measurement state of the system that Alice measured, in analogy to the quantum case (Figure 4). The crucial detail is that Ursula is not allowed to open her box (in green) and access the circuitry inside. Note that there are other possibilities for modelling measurements — this is the simplest one that still allows us to derive the paradox. For example, the choice of keeping two binary inputs in Ursula's box and discarding the second one (replacing it with 0) is an arbitrary one, picked for simplicity. Details and proofs in Appendix C.

a , and that Bob makes his measurement Y at time $t = 2$, obtaining outcome b . As usual, if Alice and Bob were to communicate at this point, they would find that $XY = a \oplus b$, and indeed the propositions $\phi_{X,a}$ and $\phi_{Y,b}$ that represent their subjective measurement experience would hold. But now suppose that Alice and Bob do not get the chance to communicate and compare their input and outputs; instead, at time $t = 3$, an observer Ursula, who models Alice's measurement as in Figure 9a, runs the box corresponding to Alice's half of the PR box and Alice's memory, and applies the post-processing of Figure 9b. Ursula's input is \tilde{X} and her output is \tilde{a} . Then the claim is that $\tilde{X}Y = \tilde{a} \oplus b$: that is, Ursula and Bob effectively share a PR box. This is proven in Appendix C. We now have all the ingredients needed to find a multi-agent epistemic paradox in box world.

4 Finding the paradox

In this section we find a scenario in box world where reasoning, physical agents reach a logical paradox. We compare it to the result to the contradiction obtained by Frauchiger and Renner [1] in the next section.

Experimental setup. The proposed thought experiment is similar in spirit to the one proposed by Frauchiger and Renner [1] (recall Figure 1). Alice and Bob share a PR box; they each will measure their half of the PR box and store the outcomes in their local memories. Let Alice’s lab be located inside the lab of another agent, Ursula’s lab such that Ursula can now perform joint measurements on Alice’s system (her half of the PR box) and memory, as seen in the previous section. Similarly, let Bob’s lab be located inside Wigner’s lab, such that Wigner can perform joint measurements on Bob’s system and memory. We assume that Alice’s and Bob’s labs are isolated such that no information about their measurement outcomes leaks out. The protocol, as shown in Figure 1b, is the following:

- t=1** Alice measures her half of the PR box, with measurement setting X , and stores the outcome a in her memory A .
- t=2** Bob measures his half of the PR box, with measurement setting Y , and stores the outcome b in his memory B .
- t=3** Ursula measures the box corresponding to Alice’s lab (as in Figure 9b), with measurement setting $\tilde{X} = X \oplus 1$, obtaining outcome \tilde{a} .
- t=4** Wigner measures the box corresponding to Bob’s lab, with measurement setting $\tilde{Y} = Y \oplus 1$, obtaining outcome \tilde{b} .

Agents can agree on their measurement settings beforehand, but should not communicate once the experiment begins. The trust relation, which specifies which agents consider each other to be rational agents (as opposed to mere physical systems), is

$$\begin{aligned} A_{t=1,2} &\rightsquigarrow B_{t=1,2} \\ B_{t=2,3} &\rightsquigarrow U_{t=3} \\ U_{t=3,4} &\rightsquigarrow W_{t=4} \\ W_{t=4} &\rightsquigarrow A_{t=1}. \end{aligned}$$

The common knowledge \mathbb{T} shared by all four agents includes the PR box correlations, the way the external agents model Alice and Bob’s measurements, and the trust structure above.

Reasoning. Now the agents can reason about the events in other agents’ labs. We take here the example where the measurement settings are $X = Y = 0$, $\tilde{X} = \tilde{Y} = 1$, and where Wigner obtained the outcome $\tilde{b} = 0$; the reasoning is analogous for the remaining cases.

1. *Wigner reasons about Ursula’s outcome.* At time $t = 4$, Wigner knows that, by virtue of their information-preserving modelling of Alice and Bob’s measurements, he and Ursula effectively shared a PR box ⁷. He can therefore use the PR correlations $\tilde{X}\tilde{Y} = \tilde{a} \oplus \tilde{b}$ to conclude that Ursula’s output must be 1,

$$K_W(\tilde{b} = 0 \implies \tilde{a} = 1).$$

⁷See Appendix C for a proof.

2. *Wigner reasons about Ursula's reasoning.* Now Wigner thinks about what Ursula may have concluded regarding Bob's outcome. He knows that at time $t = 3$, Ursula and Bob effectively shared a PR box⁷, satisfying $\tilde{X}Y = \tilde{a} \oplus b$, and that therefore Ursula must have concluded

$$K_W K_U (\tilde{a} = 1 \implies b = 1).$$

3. *Wigner reasons about Ursula's reasoning about Bob's reasoning.* Next, Wigner wonders "What could Ursula, at time $t = 3$, conclude about Bob's reasoning at time $t = 2$?" Well, Wigner knows that she knows that Bob knew that at time $t = 2$ he effectively shared a PR box with Alice, satisfying $XY = a \oplus b$, and therefore concludes

$$K_W K_U K_B (b = 1 \implies a = 1).$$

4. *Wigner reasons about Ursula's reasoning about Bob's reasoning about Alice's reasoning.* We are almost there. Now Wigner thinks about Alice's perspective at time $t = 1$, through the lenses of Bob (at time $t = 2$) and Ursula ($t = 3$). Back then, Alice knew that she obtained some outcome a , and that Wigner would model Bob's measurement in an information-preserving way, such that Alice (at time $t = 1$) and Wigner (of time $t = 4$) share an effective PR box⁷, satisfying $X\tilde{Y} = a \oplus \tilde{b}$, which results, in particular, in

$$K_W K_U K_B K_A (a = 1 \implies \tilde{b} = 1).$$

5. *Wigner applies trust relations.* In order to combine the statements obtained above, we need to apply the trust relations described above, starting from the inside of each proposition, for example,

$$\begin{aligned} & K_W K_U K_B K_A (a = 1 \implies \tilde{b} = 1) \\ \implies & K_W K_U K_B (a = 1 \implies \tilde{b} = 1) & [A \rightsquigarrow B] \\ \implies & K_W K_U (a = 1 \implies \tilde{b} = 1) & [B \rightsquigarrow U] \\ \implies & K_W (a = 1 \implies \tilde{b} = 1), & [U \rightsquigarrow W] \end{aligned}$$

and similarly for the other statements, so that we obtain

$$\begin{aligned} & K_W [(\tilde{b} = 0 \implies \tilde{a} = 1) \wedge (\tilde{a} = 1 \implies b = 1) \wedge (b = 1 \implies a = 1) \wedge (a = 1 \implies \tilde{b} = 1)] \\ \implies & K_W (\tilde{b} = 0 \implies \tilde{b} = 1). \end{aligned}$$

We could have equally taken the point of view of any other observer, and from any particular outcome or choice of measurement, and through similar reasoning chains reached the following contradictions,

$$\begin{aligned} & K_A [(a = 0 \implies a = 1) \wedge (a = 1 \implies a = 0)], \\ & K_B [(b = 0 \implies b = 1) \wedge (b = 1 \implies b = 0)], \\ & K_U [(\tilde{a} = 0 \implies \tilde{a} = 1) \wedge (\tilde{a} = 1 \implies \tilde{a} = 0)], \\ & K_W [(\tilde{b} = 0 \implies \tilde{b} = 1) \wedge (\tilde{b} = 1 \implies \tilde{b} = 0)]. \end{aligned}$$

5 Discussion

We have generalized the conditions of the Frauchiger-Renner theorem and made them applicable to arbitrary physical theories, including the framework of *generalized probability theories*. We then applied these conditions to the GTP of box world and found an experimental setting that leads to a multi-agent epistemic paradox.

5.1 Comparison with the quantum thought experiment

We showed that box world agents reasoning about each others' knowledge can come to a deterministic contradiction, which is stronger than the original paradox, as it can be reached without post-selection, from the point of view of every agent and for any measurement outcome obtained by them.

Strong contextuality and post-selection. In contrast to the original Frauchiger-Renner experiment of [1], no post-selection was required to arrive at this contradictory chain of statements as, in fact, all the implications above are symmetric, for example

$$\tilde{a} = 0 \iff b = 0 \iff a = 0 \iff \tilde{b} = 0 \iff \tilde{a} = 1.$$

As a result, one can arrive at a similar (symmetric) paradoxical chain of statements irrespective of the choice of agent and outcome for the first statement. In other words, irrespective of the outcomes observed by every agent, each agent will arrive at a contradiction when they try to reason about the outcomes of other agents. This is because, as shown in [8], the PR box exhibits strong contextuality and no global assignments of outcome values for all four measurements exists for any choice of local assignments. In contrast, the original paradox of [1] admits the same distribution as that of Hardy's paradox [7]. It is shown in [8] that this distribution is an example of logical contextuality where for a particular choice of local assignments (the ones that are post-selected on in the original Frauchiger-Renner experiment), a global assignment of values compatible with the support of the distribution fails to exist, but this is not true for all local assignments. This makes the paradox even stronger in box world, since it can be found without post-selection and by any of the agents, for any outcome that they observe. In particular, the paradox would already arise in a single run of the experiment. For a simple method to enumerate all possible contradictory statements that the agents may make, see the analysis of the PR box presented in [8]. A detailed analysis of the relation between Frauchiger-Renner type paradoxes and contextuality will appear in future work.

Communication vs prepare-and-measure. One might note that a distinction between our proposal and the original Frauchiger-Renner experiment is that there is no communication between Alice and Bob in our PR box version. However, the original quantum scenario can be replaced by a protocol where Alice and Bob receive an appropriately prepared quantum state and perform measurements on it without communicating to each other, and the original paradox would still hold in such a case (Figure 1a).

5.2 Physical measurements in box world

Since we lack a physical theory to explain how measurements and transformations are instantiated for generalised non-signalling boxes, and only have access to their input/output behaviour, all allowed transformations consist of pre- and post-processing. In the quantum case, we have in addition to a description of possible input-output correlations, a mathematical framework for the underlying states producing those correlations, the theory of von-Neumann measurements and transformations as CPTP maps. As we cannot "open" the boxes in GPTs in a same way we are allowed to do so in quantum theory, introduction of dynamical features (for example, a memory update algorithm) in box world is less intuitive and requires additional constructions. One way of doing so was introduced in [17]: there authors restrict the set of allowed PR boxes; then it is possible to come up with a coupler device which implements swapping and teleportation, as well as recover quantum correlations for this particular set of states. In the following, we elaborate on these peculiar features of box world while comparing them with the usual quantum setting.

The *outside* perspective: measurements vs wirings. When a box world experimenter (say, Alice) measures a box (say a gbit characterised by 2 fiducial measurements, $X = 0, X = 1$), the box which is a one-time function essentially disappears (from her perspective) after the measurement, leaving her only with the classical measurement outcome characterised by a single fiducial measurement (say $X = 0$). If Alice, given just a single copy of this box (a gbit), measured $X = 0$ on it obtaining the classical outcome a , she can no longer perform the $X = 1$ on her system since the box was used up in the first measurement. She could however reprepare a new box using the information she gained from this measurement⁸. In contrast, a qubit measured in the Z basis is classical in this basis post-measurement but it can still be measured in a different basis, the system does not disappear since we have a physical theory of how measurements are instantiated for quantum boxes.

For an outside observer, Ursula who models Alice and her memory as a box world system, the box does not disappear until she (Ursula) measures and it is possible for Ursula to measure $X = 1$ on the joint system of Alice's box and her memory even after Alice has measured $X = 0$ on her box, as long as Ursula does not know Alice's outcome⁹. However, there is no way for Ursula to tell which measurement Alice performed or whether she measured at all. Alice could have simply wired the boxes as in Figure 9a without actually performing the measurement, and Ursula will not know the difference. The joint state of Alice's system S and her memory M that Ursula sees in all 3 cases, 1) Alice simply wired the system and her memory as in Figure 9a, 2) she performed the $X = 0$ measurement, 3) she performed the $X = 1$ measurement, will be the same as that of Equation 13 (or equivalently Equation 14). From the outside perspective, measurements are equivalent to classical wirings and Alice's memory only gets classically correlated with her outcome (if she had measured) from Ursula's perspective. This is in stark contrast with the quantum case where Alice's memory gets entangled with the system post-measurement and the post-measurement state depends on the basis in which Alice measured her system. For example, if Alice's qubit S starts off in the normalised pure state $|\psi\rangle = \alpha|0\rangle_S + \beta|1\rangle_S$ and her memory M initialised to $|0\rangle_M$, the initial state of her system and memory from Ursula's perspective is

$$|\Psi\rangle_{SM}^{in} = [\alpha|0\rangle_S + \beta|1\rangle_S] \otimes |0\rangle_M = \left[\left(\frac{\alpha + \beta}{\sqrt{2}} \right) |+\rangle_S + \left(\frac{\alpha - \beta}{\sqrt{2}} \right) |-\rangle_S \right] \otimes |0\rangle_M. \quad (7)$$

If Alice then measures the system in the computational (Z) basis, the post-measurement state from Ursula's perspective is

$$|\Psi\rangle_{SM}^{out,Z} = \alpha|0\rangle_S|0\rangle_M + \beta|1\rangle_S|1\rangle_M, \quad (8)$$

which is an entangled state. If instead, Alice measured in the Hadamard (X) basis, the post-measurement state would be $|\Psi\rangle_{SM}^{out,X} = \left(\frac{\alpha + \beta}{\sqrt{2}} \right) |+\rangle_S |0\rangle_M + \left(\frac{\alpha - \beta}{\sqrt{2}} \right) |-\rangle_S |1\rangle_M$ or equivalently,

$$|\Psi\rangle_{SM}^{out,X} = \left(\frac{\alpha + \beta}{2} \right) |00\rangle_{SM} + \left(\frac{\alpha + \beta}{2} \right) |10\rangle_{SM} + \left(\frac{\alpha - \beta}{2} \right) |01\rangle_{SM} - \left(\frac{\alpha - \beta}{2} \right) |11\rangle_{SM}. \quad (9)$$

Clearly the measurement statistics of $|\Psi\rangle_{SM}^{in}$, $|\Psi\rangle_{SM}^{out,Z}$ and $|\Psi\rangle_{SM}^{out,X}$ are different and Ursula can thus (in principle) tell whether or not Alice performed a measurement and which measurement was performed by her.

Supergluing of non-signalling boxes. For the memory update circuit (from Ursula's perspective) of Figure 9a, and the initial state of Equation 12, the final state would be $\mathbf{P}_{fin}^{SM} =$

⁸See Appendix D for further details.

⁹If Ursula learns Alice's outcome, she takes Alice's perspective and the box disappears for her as well.

$(p \ 0 \ 0 \ 1-p|p \ 0 \ 0 \ 1-p|q \ 0 \ 0 \ 1-q|q \ 0 \ 0 \ 1-q)^T_{SM}$. Note that while the reduced final state of S does not depend on the input X' to M , the reduced final state on Alice's memory M , \mathbf{P}_{fin}^M clearly depends on the input X of the system S if $p \neq q$. If $X = 0$, $\mathbf{P}_{fin}^M = (p \ 1-p|p \ 1-p)^T$ and if $X = 1$, $\mathbf{P}_{fin}^M = (q \ 1-q|q \ 1-q)^T$, i.e., the systems S and M are *signalling*. This is expected since there is clearly a transfer of information from S to M during the measurement as seen in Figure 6. However, this means that the state \mathbf{P}_{fin}^{SM} is not a valid box world state of 2 systems S and M but a valid state of a single system SM i.e., after Alice performs her wiring/measurement, it is not possible to physically separate Alice's system S from her memory M from Ursula's perspective. For if this were possible, there would be a violation of the no-signalling principle and the notion of relativistic causality. In quantum theory, on the other hand it is always possible to perform separate measurements on Alice's system and her memory even after she measures. We call this feature *supergluing* of post-measurement boxes, where it is no longer possible for Ursula to separately measure S or M , but she can only jointly measure SM as though it were a single system. Note that this is only the case for $p \neq q$ and in our example with the PR-box (Section 4), $p = q = 1/2$ and \mathbf{P}_{fin}^{SM} remains a valid bipartite non-signalling state in this particular, fine-tuned case of the PR box.

These features of the memory update in box world are certainly not desirable ones and not what one would expect to find in a physical theory. These give further intuition for why PR boxes have not yet been found in nature which appears to obey the laws of quantum physics. One of the main contributions of this paper is the finding that despite these peculiar features of box world and the fact that it has no entangling bipartite joint measurements (a crucial step in the original quantum paradox), a consistent outside perspective of the memory update exists such that with our generalised assumptions, an FR type paradox can be recovered. This indicates that non-locality and not unitarity might be the main reason for FR type paradoxes.

5.3 Characterization of general theories

While we have shown that a consistency paradox, similar to the one arising in the Frauchiger-Renner setup, can also be adapted for the box world in terms of GPTs, it still remains unclear how to characterize all possible theories where it is possible to find a setup leading to a contradiction. Essentially, one has to restrict the class of such theories and identify the properties of these theories that make such paradoxes possible. It seems that there is an apparent link to the notion of contextuality, yet not all types of contextuality admit logical consistency paradoxes similar to Frauchiger-Renner; we will investigate these connections further in our future work.

Additionally, it is still an open problem to find an operational way to state the outside view of measurements (and a memory update operation), for theories without a prior notion of subsystems and a tensor rule for composing them. This will allow us to search for multi-agent logical paradoxes in field theories, for example. One possible direction is to use notions of effective locality, as outlined for example in [12].

Acknowledgements

We thank Roger Colbeck, Matt Leifer, Sandu Popescu and Renato Renner for valuable discussions. VV acknowledges support from the Department of Mathematics, University of York. NN and LdR acknowledge support from the Swiss National Science Foundation through SNSF project No. 200020_165843 and through the the National Centre of Competence in Research *Quantum Science and Technology* (QSIT). LdR further acknowledges support from the FQXi grant *Physics of the observer*.

APPENDIX

A Modal logic

Here we shortly sum up the important features of modal logic. Importantly, modal logic applies to most classical multi-agent setups, and simply provides a compact mathematical way to capture the intuitive laws commonly used for reasoning.

A.1 Kripke structures

In modal logic, a set Σ of possible states (or alternatives, or *worlds*) is introduced [18]: for example, in a world s_1 the key value is $k = 1$ and Eve does not know it, and in a state s_2 Eve could know that $k = 0$. The truth value of a proposition ϕ is then assigned depending on the possible world in Σ , and can differ from one possible world to another. In order to formalize the simple rules agents use for reasoning, we will first provide a structure which serves as a complete picture of the setup the agents are in, and then discuss the elements of the structure.

Definition 8 (Kripke structure) *A Kripke structure M for n agents over a set of statements Φ is a tuple $\langle \Sigma, \pi, \mathcal{K}_1, \dots, \mathcal{K}_n \rangle$ where Σ is a non-empty set of states, or possible worlds, π is an interpretation, and \mathcal{K}_i is a binary relation on Σ .*

The interpretation π is a map $\pi : \Sigma \times \Phi \rightarrow \{\mathbf{true}, \mathbf{false}\}$, which defines a truth value of a statement $\phi \in \Phi$ in a possible world $s \in \Sigma$.

\mathcal{K}_i is a binary equivalence relation on a set of states Σ , where $(s, t) \in \mathcal{K}_i$ if agent i considers world t possible given his information in the world s .

The truth assignment tells us if the proposition $\phi \in \Phi$ is true or false in a possible world $s \in \Sigma$; for example, if $\phi =$ ‘‘Alice has a secret key,’’ and s is a world where there is an individual named Alice who indeed possesses a secret key, then $\pi(s, \phi) = \mathbf{true}$. The truth value of a statement in a given structure M might vary from one possible world to another; we will denote that ϕ is true in world s of a structure M by $(M, s) \models \phi$, and $\models \phi$ will mean that ϕ is true in any world s of a structure M .

A.2 Axioms of knowledge (weak version)

In order to operate the statements agents produce, we have to establish certain rules which are used to compress or judge the statements. These are the axioms of knowledge [19]. They might seem trivial in the light of our everyday reasoning, yet given our awareness of the quantum case, we will treat them carefully. Here we present the reader with a weaker version of the axioms (which includes Trust axiom) that we have developed in previous work [5].

Distribution axiom allows agents combine statement which contain inferences:

Axiom 1 (Distribution axiom.) *If an agent is aware of a fact ϕ and that a fact ψ follows from ϕ , then the agent can conclude that ψ holds:*

$$(M, s) \models (K_i \phi \wedge K_i(\phi \Rightarrow \psi)) \Rightarrow (M, s) \models K_i \psi.$$

Knowledge generalization rule permits agents use commonly shared knowledge:

Axiom 2 (Knowledge generalization rule.) *All agents know all the propositions that are valid in a structure:*

$$\text{if } (M, s) \models \phi \ \forall s \text{ then } \models K_i \phi \ \forall i.$$

Positive and negative introspection axioms highlight the ability of an agent to reflect upon her knowledge:

Axiom 3 (Positive and negative introspection axioms.) *Agents can perform introspection regarding their knowledge:*

$$\begin{aligned} (M, s) \models K_i \phi &\Rightarrow (M, s) \models K_i K_i \phi \text{ (Positive Introspection),} \\ (M, s) \models \neg K_i \phi &\Rightarrow (M, s) \models K_i \neg K_i \phi \text{ (Negative Introspection).} \end{aligned}$$

We also equip the logical skeleton of the setting with so-called trust structure, which governs the way the information is passed on between agents:

Definition 9 (Trust) *We say that an agent i trusts an agent j (and denote it by $j \rightsquigarrow i$) if and only if*

$$K_i K_j \phi \implies K_i \phi,$$

for all ϕ .

In the Frauchiger-Renner setup, as well as in the thought experiment presented in this paper, we consider the following trust structure between agents:

$$A \rightsquigarrow B \rightsquigarrow U \rightsquigarrow W \rightsquigarrow A. \quad (10)$$

Further discussion on axioms of modal logic and their application in quantum mechanics can be found in our paper [5].

B Generalized probabilistic theories

In quantum theory, systems are described by states that live in a Hilbert space, measurements and transformations on these states are represented by CPTP maps and the Born rule specifies how to obtain the probabilities of possible measurement outcomes given these states and measurements. In more general theories, there is no reason to assume Hilbert spaces or CPTP maps. In fact such a description of the state space and operations may not even be available, systems may be described as black boxes taking in classical inputs (choice of measurements) and giving classical outputs (measurement outcomes). What we can demand is that the theory provides a way for agents to predict the probabilities of obtaining various outputs based on their input choice and some operational description of the box.

Barrett derived the mathematical structure of the state-space of composite systems and allowed operations on systems from a few reasonable, physically motivated assumptions [3]. We follow his formalism here. Later, Gross *et al.* found restrictions on the reversible dynamics of maximally non-local GPTs [15] showing that all reversible operations on box-world are trivial i.e., they map product states to product states and cannot correlate initially uncorrelated systems. In accordance with this, our memory update procedure that maps the initial product state \mathbf{P}_{in}^{SM} (Equation 12) to the final correlated state of the system and memory \mathbf{P}_{fin}^{SM} (Equation 13 or equivalently Equation 14) is an irreversible transformation in contrast to the quantum case where the corresponding transformation is a unitary and hence reversible.

B.1 Observing outcomes

In Section 3, we briefly reviewed states and transformations in GPTs, in particular box world, here we go into further detail. Consider a GPT, \mathbb{T} . Denoting the set of all allowed states of a system in \mathbb{T} by \mathcal{S} , any valid transformation on a normalised GPT state $\mathbf{P} \in \mathcal{S}$ maps it to

another normalised GPT state in \mathcal{S} . Consequently, is linear and can be represented by a matrix M such that $\mathbf{P} \rightarrow M \cdot \mathbf{P}$ under this transformation and $M \cdot \mathbf{P} \in \mathcal{S}$ [3]. Further, operations that result in different possible outcomes can be associated with a set of transformations, one for each outcome. These also give an operational meaning to unnormalised states where $|\mathbf{P}| = \sum_i P(a = i|X = j) = c \quad \forall j, c \in [0, 1]$ (i.e., the norm is independent of the value of j). Such an operation M on a normalised initial state \mathbf{P} can be associated with a set of matrices $\{M_i\}$ such that the unnormalised state corresponding to the i^{th} outcome is $M_i \cdot \mathbf{P}$. Then the probability of obtaining this outcome is simply the norm of this unnormalised state, $|M_i \cdot \mathbf{P}|$ and the corresponding normalized final state is $M_i \cdot \mathbf{P} / |M_i \cdot \mathbf{P}|$. A set $\{M_i\}$ represents a valid operation if the following hold [3].

$$0 \leq |M_i \cdot \mathbf{P}| \leq 1 \quad \forall i, \mathbf{P} \in \mathcal{S} \quad (11a)$$

$$\sum_i |M_i \cdot \mathbf{P}| = 1 \quad \forall \mathbf{P} \in \mathcal{S} \quad (11b)$$

$$M_i \cdot \mathbf{P} \in \mathcal{S} \quad \forall i, \mathbf{P} \in \mathcal{S} \quad (11c)$$

This is the analogue of quantum Born rule for GPTs. Box world is a GPT where the state space \mathcal{S} consists of all normalized states \mathbf{P} whose entries are valid probabilities (i.e., $\in [0, 1]$) and satisfy the *no-signalling* constraints i.e., for a N -partite state \mathbf{P} , the marginal term $\sum_{a_i} P(a_1, \dots, a_i, \dots, a_N | X_1, \dots, X_i, \dots, X_N)$ is independent of the setting X_i for all $i \in \{1, \dots, N\}$ ¹⁰

When the GPT \mathbb{T} is box world, the conditions of Equations 11a-11c result in the characterization of measurements and transformations in the theory in terms of classical circuits or *wirings* as shown in [3]. It suffices for the purpose of this paper to take that characterisation as the common knowledge of agents in the theory. In the original quantum paradox [1], the Born rule is taken as common knowledge and here, the common knowledge consists of characterisations that follow from the box world analogue of the born rule (Equations 11a-11c). We summarise the results of [3] characterising allowed transformations and measurements in box world and will only consider normalization-preserving transformations.

- **Transformations:**

- *Single system:* All transformations on single box world systems are relabellings of fiducial measurements or outcomes or a convex combination thereof.
- *Bipartite system:* Let X and Y be fiducial measurements performed on the transformed bipartite system with corresponding outcomes a and b , then all transformations of 2-gbit systems can be decomposed into convex combinations of classical circuits of the following form: A fiducial measurement $X' = f_1(X, Y)$ is performed on the initial state of the first gbit resulting in the outcome a' followed by a fiducial measurement $Y' = f_2(X, Y, X')$ on the initial state of the second gbit resulting in the outcome b' . The final outcomes are given as $(a, b) = f_3(X, Y, a', b')$, where f_1, f_2 and f_3 are arbitrary functions.

- **Measurements:**

- *Single system:* All measurements on single box world systems are either fiducial measurements with outcomes relabelled or convex combinations of such.
- *Bipartite system:* All bipartite measurements on 2-gbit systems can be decomposed into convex combinations of classical circuits of the following form (Figure 6): A fiducial measurement X is performed on the initial state of the first gbit resulting

¹⁰This is in the spirit of relativistic causality since one would certainly expect that the input of one party does not affect the output of others when they are all space-like separated from each other.

in the outcome a' followed by a fiducial measurement $Y = f(a')$ on the second gbit resulting in the outcome b' . The final outcome is $a = f'(a', b')$, where f and f' are arbitrary functions.

Remark: Note that an agent (Alice) who measures a box world system only sees a classical final state that is the classical measurement outcome since the box is a one-time input/output function. Alice could use Equations 11a-11c to calculate the probabilities of obtaining different outcomes given the measurement she performs and prepare a new box (a new input/output function) depending on the measurement and outcome she just obtained (and has stored in her memory). An outside agent (Ursula) who does not know Alice's measurement outcome would see correlations between Alice's system and memory and would describe the measurement by an irreversible transformation, more specifically a classical wiring between Alice's system and memory as shown in the following Section.

C Memory update in box world (proofs)

C.1 Single lab

In this section, we describe how a box world agent would measure a system and store the result in a memory. From the perspective of an outside observer (who does not know the outcome of the agent's measurement), we describe the initial and final states of the system and memory before and after the measurement as well as the transformation that implements this memory update in box world. In the quantum case, any initial state of the system S is mapped to an isomorphic joint state of the system S and memory M (see Equation 1) and hence the memory update map that maps the former to the latter (an isometry in this case¹¹) satisfies Definition 6 of an information-preserving memory update. We will now characterise the analogous memory update map in box world and show that it also satisfies Definition 6. In the following, we consider gbit systems for simplicity but the argument holds for larger dimensional box world systems as well.

Theorem 10 *In box world, there exists a valid transformation u that maps every arbitrary, normalized gbit state \mathbf{P}_{in}^S on the system S to an isomorphic final state \mathbf{P}_{fin}^{SM} on the system S and memory M (also a gbit) and hence constitutes an information-preserving memory update (Definition 6).*

Proof: We start with the system in an arbitrary, normalized gbit state $\mathbf{P}_{in}^S = (p \ 1-p|q \ 1-q)^T$ (where the subscript T denotes transpose and $p, q \in [0, 1]$) and the memory initialised to one of the 4 pure states¹², say $\mathbf{P}_{in}^M = \mathbf{P}_1 = (1 \ 0|1 \ 0)^T$. Then the joint initial state, $\mathbf{P}_{in}^{SM} = (p \ 1-p|q \ 1-q)_S^T \otimes (1 \ 0|1 \ 0)_M^T$ of the system and memory can be written as follows, where $P_{in}(a = i, a' = j|X = k, X' = l)$ denotes the probability of obtaining the outcomes $a = i$ and $a' = j$ when performing the fiducial measurements $X = k$ and $X' = l$ on the system and

¹¹An isometry since it involves introducing an initial pure state on M and then performing a joint unitary on SM .

¹²It does not matter which pure state the memory is initialized in, a similar argument applies in all cases.

memory respectively, in the initial state \mathbf{P}_{in}^{SM} .

$$\mathbf{P}_{in}^{SM} = \left(\begin{array}{c} P_{in}(a=0, a'=0|X=0, X'=0) \\ P_{in}(a=0, a'=1|X=0, X'=0) \\ P_{in}(a=1, a'=0|X=0, X'=0) \\ P_{in}(a=1, a'=1|X=0, X'=0) \\ \hline P_{in}(a=0, a'=0|X=0, X'=1) \\ P_{in}(a=0, a'=1|X=0, X'=1) \\ P_{in}(a=1, a'=0|X=0, X'=1) \\ P_{in}(a=1, a'=1|X=0, X'=1) \\ \hline P_{in}(a=0, a'=0|X=1, X'=0) \\ P_{in}(a=0, a'=1|X=1, X'=0) \\ P_{in}(a=1, a'=0|X=1, X'=0) \\ P_{in}(a=1, a'=1|X=1, X'=0) \\ \hline P_{in}(a=0, a'=0|X=1, X'=1) \\ P_{in}(a=0, a'=1|X=1, X'=1) \\ P_{in}(a=1, a'=0|X=1, X'=1) \\ P_{in}(a=1, a'=1|X=1, X'=1) \end{array} \right)_{SM} = \left(\begin{array}{c} p \\ 0 \\ 1-p \\ 0 \\ \hline p \\ 0 \\ 1-p \\ 0 \\ \hline q \\ 0 \\ 1-q \\ 0 \\ \hline q \\ 0 \\ 1-q \\ 0 \end{array} \right)_{SM} \quad (12)$$

The rest of the proof proceeds as follows: we first describe a final state \mathbf{P}_{fin}^{SM} of the system and memory and a corresponding memory update map u that satisfy Definition 6 of a generalized information-preserving memory update. Then, we show that this map is an allowed box world transformation which completes the proof.

If an agent performs a measurement on the system, the state of the memory must be updated depending on the outcome and the final state of the system and memory after the measurement must hence be a correlated (i.e., a non-product) state. Although the full state space of the 2 gbit system SM is characterised by the 4 fiducial measurements $(X, X') \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$, Definition 6 allows us to restrict possible final states to a useful subspace of this state space that contain correlated states of a certain form. The definition requires that for every map \mathcal{E}_S on the system before measurement, there exists a corresponding map \mathcal{E}_{SM} on the system and memory after the measurement that is operationally identical. Thus it suffices if the joint final state \mathbf{P}_{fin}^{SM} belongs to a subspace of the 2 gbit state space for which only 2 of the 4 fiducial measurements are relevant for characterising the state, namely any 2 fiducial measurements on \mathbf{P}_{fin}^{SM} that are isomorphic to the 2 fiducial measurements on \mathbf{P}_{in}^S . Note that by definition of fiducial measurements, the outcome probabilities of any measurement can be found given the outcome probabilities of all the fiducial measurements and without loss of generality, we will only consider the case where the agents perform fiducial measurements on their systems.

A natural isomorphism between fiducial measurements on \mathbf{P}_{in}^S and those on \mathbf{P}_{fin}^{SM} to consider here (in analogy with the quantum case) is: $X = i \Leftrightarrow (X, X') = (i, i)$, $\forall i \in \{0, 1\}$ i.e., only consider the cases where the fiducial measurements performed on S and M are the same. Now, in order for the states to be isomorphic or operationally equivalent, one requires that performing the fiducial measurements $(X, X') = (i, i)$ on \mathbf{P}_{fin}^{SM} should give the same outcome statistics as measuring $X = 0$ on \mathbf{P}_{in}^S . This can be satisfied through an identical isomorphism on the outcomes: $a = i \Leftrightarrow (a, a') = (i, i)$, $\forall i \in \{0, 1\}$. Then the final state of the system and memory,

\mathbf{P}_{fin}^{SM} will be of the form

$$\mathbf{P}_{fin}^{SM} = \left(\begin{array}{c} P_{fin}(a = 0, a' = 0|X = 0, X' = 0) \\ P_{fin}(a = 0, a' = 1|X = 0, X' = 0) \\ P_{fin}(a = 1, a' = 0|X = 0, X' = 0) \\ P_{fin}(a = 1, a' = 1|X = 0, X' = 0) \\ \hline P_{fin}(a = 0, a' = 0|X = 0, X' = 1) \\ P_{fin}(a = 0, a' = 1|X = 0, X' = 1) \\ P_{fin}(a = 1, a' = 0|X = 0, X' = 1) \\ P_{fin}(a = 1, a' = 1|X = 0, X' = 1) \\ \hline P_{fin}(a = 0, a' = 0|X = 1, X' = 0) \\ P_{fin}(a = 0, a' = 1|X = 1, X' = 0) \\ P_{fin}(a = 1, a' = 0|X = 1, X' = 0) \\ P_{fin}(a = 1, a' = 1|X = 1, X' = 0) \\ \hline P_{fin}(a = 0, a' = 0|X = 1, X' = 1) \\ P_{fin}(a = 0, a' = 1|X = 1, X' = 1) \\ P_{fin}(a = 1, a' = 0|X = 1, X' = 1) \\ P_{fin}(a = 1, a' = 1|X = 1, X' = 1) \end{array} \right)_{SM} = \left(\begin{array}{c} p \\ 0 \\ 0 \\ 1 - p \\ \hline * \\ * \\ * \\ * \\ \hline * \\ * \\ * \\ * \\ \hline q \\ 0 \\ 0 \\ 1 - q \end{array} \right)_{SM}, \quad (13)$$

where $*$ are arbitrary, normalised entries and where $P_{fin}(a = i, a' = j|X = k, X' = l)$ denotes the probability of obtaining the outcomes $a = i$ and $a' = j$ when performing the fiducial measurements $X = k$ and $X' = l$ on the system and memory respectively, in the final state \mathbf{P}_{fin}^{SM} . This final state can be compressed since the only relevant and non-zero probabilities in \mathbf{P}_{fin}^{SM} occur when $X = X'$ and $a = a'$. We can then define new variables \tilde{X} and \tilde{a} such that $X = X' = i \Leftrightarrow \tilde{X} = i$ and $a = a' = j \Leftrightarrow \tilde{a} = j$ for $i, j \in \{0, 1\}$ and \mathbf{P}_{fin}^{SM} can equivalently be written as in Equation 14 which is clearly of the same form as \mathbf{P}_{in}^S .

$$\mathbf{P}_{fin}^{SM} \equiv \left(\begin{array}{c} P(\tilde{a} = 0|\tilde{X} = 0) \\ P(\tilde{a} = 1|\tilde{X} = 0) \\ \hline P(\tilde{a} = 0|\tilde{X} = 1) \\ P(\tilde{a} = 1|\tilde{X} = 1) \end{array} \right)_{SM} = \left(\begin{array}{c} p \\ 1 - p \\ \hline q \\ 1 - q \end{array} \right)_{SM} \quad (14)$$

Hence the initial state of the system, $\mathbf{P}_{in}^S = (p \ 1 - p|q \ 1 - q)^T$ (which is an arbitrary gbit state) is isomorphic to the final state of the system and memory, \mathbf{P}_{fin}^{SM} (as evident from Equation 14) with the same outcome probabilities for $X = 0, 1$ and $\tilde{X} = 0, 1$. This implies that for every transformation \mathcal{E}_S on the former, there exists a transformation \mathcal{E}_{SM} on the latter such that for all outside agents A_j and for all $p, q \in [0, 1]$ (i.e., all possible input gbit states on the system), $K_j\phi[\mathcal{E}_S(\mathbf{P}_{in}^S)] \Rightarrow K_j\phi[\mathcal{E}_{SM} \circ \mathbf{P}_{SM}^S M_{fin}]$, where $\mathbf{P}_{SM}^{fin} = u(\mathbf{P}_{in}^S)$. Thus any map u that maps \mathbf{P}_{in}^S to \mathbf{P}_{SM}^{fin} satisfies Definition 6.

We now find a valid box world transformation that maps the initial state \mathbf{P}_{in}^{SM} (Equation 12) to any final state of the form \mathbf{P}_{fin}^{SM} (Equation 13). This fully characterises the memory update map $u : \mathbf{P}_{in}^S \rightarrow \mathbf{P}_{SM}^{fin}$ since \mathbf{P}_{SM}^{SM} is obtained from \mathbf{P}_{in}^S by simply tensoring a pure state $(1 \ 0|1 \ 0)_M^T$.

Noting that all bipartite transformations in box world can be decomposed to a classical circuit, In Figure 10, we construct an explicit circuit of this form that converts \mathbf{P}_{in}^{SM} to \mathbf{P}_{fin}^{SM} . By construction, we only need to consider the case of $X = X'$ since for $X \neq X'$, the entries of \mathbf{P}_{fin}^{SM} can be arbitrary and are irrelevant to the argument. For $X \neq X'$, one can consider any such

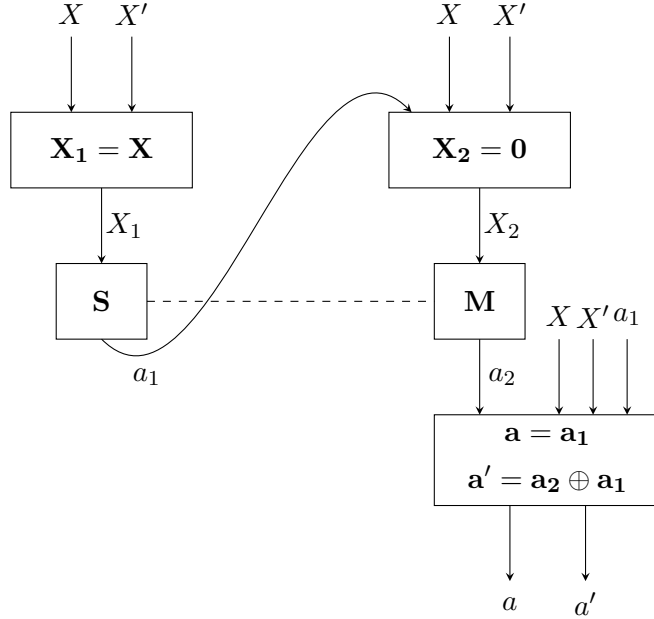


Figure 10: **Classical circuit decomposition of memory update transformation \mathcal{T}** : Let X and X' be the fiducial measurements on the transformed system with outcomes a , a' and \mathcal{T} be the transformation that maps the initial state \mathbf{P}_{in}^{SM} to a final state \mathbf{P}_{fin}^{SM} . Noting that we only need to consider the case of $X = X'$ since for $X \neq X'$, the entries of \mathbf{P}_{fin}^{SM} are arbitrary, this transformation can be equivalently obtained by the circuit shown here i.e., 1) Choose $X_1 = X (= X')$ and perform this fiducial measurement on the initial state of the system \mathbf{P}_{in}^S to obtain the outcome a_1 . 2) Fix $X_2 = 0$ (or $X_2 = 1$) and perform this fiducial measurement on the initial state of the memory \mathbf{P}_{in}^M (initialized to the pure state $(1 \ 0 | 1 \ 0)_M^T$) to obtain the outcome a_2 . 3) Set $a = a_1$. 4) If $a_1 = 1$, set $a' = a_2$, otherwise set $a' = a_2 \oplus 1$, where \oplus denotes modulo 2 addition.

circuit description and it is easy to see that $\mathbf{P}_{in}^{SM} = (p \ 1-p | q \ 1-q)_S^T \otimes (1 \ 0 | 1 \ 0)_M^T$ is indeed transformed into $\mathbf{P}_{fin}^{SM} = (p \ 0 \ 0 \ 1-p | * \ * \ * \ *)_S^T \otimes (* \ * \ * \ *)_M^T$ through the transformation \mathcal{T} defined by these sequence of steps. For example, if the circuit description for the $X \neq X'$ case is same as that for the $X = X'$ case, then we get the memory update map of Figure 9a which corresponds to performing a fixed measurement $X' = 0$ on the initial state of M and a classical CNOT on the output wire of M controlled by the output wire of S ¹³. The final state in that case is $(p \ 0 \ 0 \ 1-p | p \ 0 \ 0 \ 1-p | q \ 0 \ 0 \ 1-q | q \ 0 \ 0 \ 1-q)_S^T$. Note that the memory update transformation $\mathcal{T} : \mathbf{P}_{in}^{SM} \rightarrow \mathbf{P}_{fin}^{SM}$ and hence the resulting map u are not reversible. This is expected since the initial state \mathbf{P}_{in}^{SM} is a product state while the final state \mathbf{P}_{fin}^{SM} is clearly not (since S and M are correlated for an outside observer), and [15] shows that all reversible transformations in box world map product states to product states. \square

C.2 Two labs sharing initial correlations

So far, we have considered a single agent measuring a system in her lab. We can also consider situations where multiple agents jointly share a state and measure their local parts of the state, updating their corresponding memories. One might wonder whether the initial correlations in

¹³The output wires of boxes carry classical information after the measurement.

the shared state are preserved once the agents measure it to update their memories (clearly the local measurement probabilities remain unaltered as we saw in this section). The answer is affirmative and this is what allows us to formulate the Frauchiger-Renner paradox in box world as done in the Section 4, even though a coherent copy analogous to the quantum case does not exist here.

Suppose that Alice and Bob share the PR box correlations $X.Y = a \oplus b$ (where \cdot denotes multiplication and \oplus , bitwise addition), where X and Y are fiducial measurements on Alice and Bob's gbits A and B , and a, b are the corresponding outcomes. The associated box world state vector is as follows.

$$\mathbf{P}_{in}^{AB} = \begin{pmatrix} P(a=0, b=0|X=0, Y=0) \\ P(a=0, b=1|X=0, Y=0) \\ P(a=1, b=0|X=0, Y=0) \\ P(a=1, b=1|X=0, Y=0) \\ \hline P(a=0, b=0|X=0, Y=1) \\ P(a=0, b=1|X=0, Y=1) \\ P(a=1, b=0|X=0, Y=1) \\ P(a=1, b=1|X=0, Y=1) \\ \hline P(a=0, b=0|X=1, Y=0) \\ P(a=0, b=1|X=1, Y=0) \\ P(a=1, b=0|X=1, Y=0) \\ P(a=1, b=1|X=1, Y=0) \\ \hline P(a=0, b=0|X=1, Y=1) \\ P(a=0, b=1|X=1, Y=1) \\ P(a=1, b=0|X=1, Y=1) \\ P(a=1, b=1|X=1, Y=1) \end{pmatrix}_{AB} = \begin{pmatrix} \frac{1}{2} \\ 0 \\ 0 \\ \frac{1}{2} \\ \hline \frac{1}{2} \\ 0 \\ 0 \\ \frac{1}{2} \\ \hline \frac{1}{2} \\ 0 \\ 0 \\ \frac{1}{2} \\ \hline 0 \\ \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{pmatrix}_{AB} \quad (15)$$

Theorem 11 *Let Alice and Bob share the PR box state \mathbf{P}_{in}^{AB} representing the correlations $X.Y = a \oplus b$, locally perform a fiducial measurement on their half of the state and store the outcome in their local memories A' and B' . Then the final joint state $\mathbf{P}_{fin}^{\tilde{A}\tilde{B}}$ of the joint systems $\tilde{A} := AA'$ and $\tilde{B} := BB'$ as described by outside agents is an identical PR state between the \tilde{A} and \tilde{B} i.e., local memory updates by Alice and Bob preserves the PR box correlations initially shared between them.*

Proof: Let $X', a' \in \{0, 1\}$ and $Y', b' \in \{0, 1\}$ be the fiducial measurements and outcomes for the memory systems A' and B' (also gbits) respectively. We describe the measurement and memory update process for each agent separately and characterise the final state of Alice's and Bob's systems and memories after the process as would appear to outside agents who do not have access to Alice and Bob's measurement outcomes. This analysis does not depend on the order in which Alice and Bob perform the measurement as the correlations are symmetric between them, so without loss of generality, we can consider Alice's measurement first and then Bob's.

Suppose that Alice's memory A' is initialised to the state $\mathbf{P}_{in}^{A'} = \mathbf{P}_1 = (1 \ 0|1 \ 0)_A^T$. Then the joint initial state of the Alice's and Bob's system and Alice's memory as described by an agent Ursula outside Alice's lab is $\mathbf{P}_{in}^{ABA'} = (\mathbf{P}_{in}^{PR})^{AB} \otimes (\mathbf{P}_1)^{A'}$ which can be expanded as follows where a block containing " stands for "this block has the same entries as the previous block" and $P_{in}(a=i, b=j, a'=k|X=l, Y=m, X'=n)$ represents the probability of obtaining the outcomes $a=i, b=j, a'=k$ when performing the fiducial measurements $X=l, Y=m, X'=n$ on the initial state $\mathbf{P}_{in}^{ABA'}$.

$$\mathbf{P}_{in}^{ABA'} = \left(\begin{array}{c} P_{in}(a = 0, b = 0, a' = 0 | X = 0, Y = 0, X' = 0) \\ P_{in}(a = 0, b = 0, a' = 1 | X = 0, Y = 0, X' = 0) \\ P_{in}(a = 0, b = 1, a' = 0 | X = 0, Y = 0, X' = 0) \\ P_{in}(a = 0, b = 1, a' = 1 | X = 0, Y = 0, X' = 0) \\ P_{in}(a = 1, b = 0, a' = 0 | X = 0, Y = 0, X' = 0) \\ P_{in}(a = 1, b = 0, a' = 1 | X = 0, Y = 0, X' = 0) \\ P_{in}(a = 1, b = 1, a' = 0 | X = 0, Y = 0, X' = 0) \\ P_{in}(a = 1, b = 1, a' = 1 | X = 0, Y = 0, X' = 0) \\ \hline \cdot \\ \cdot \\ \cdot \end{array} \right)_{ABA'} = \left(\begin{array}{c} \frac{1}{2} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \frac{1}{2} \\ 0 \\ \text{"} \\ \text{"} \\ \text{"} \\ \text{"} \\ \text{"} \\ \text{"} \\ 0 \\ 0 \\ \frac{1}{2} \\ 0 \\ \frac{1}{2} \\ 0 \\ 0 \\ 0 \\ \text{"} \end{array} \right) \begin{array}{l} X = 0, Y = 0, X' = 0 \\ X = 0, Y = 0, X' = 1 \\ X = 0, Y = 1, X' = 0 \\ X = 0, Y = 1, X' = 1 \\ X = 1, Y = 0, X' = 0 \\ X = 1, Y = 0, X' = 1 \\ X = 1, Y = 1, X' = 0 \\ X = 1, Y = 1, X' = 1 \end{array}$$

Now, Ursula will describe the transformation on AA' through the memory update transformation \mathcal{T} of Figure 10. Let $\mathbf{P}_{fin}^{ABA'}$ be the final state that results by applying this map to the systems AA' in the initial state $\mathbf{P}_{in}^{ABA'}$. By the same argument as in Theorem 10, we only need to consider the cases when the same fiducial measurement is performed on Alice's system and memory i.e., only cases where $X = X'$. Then the only relevant measurement choices for describing the final state are $(X, Y, X') \in \{(0, 0, 0), (0, 1, 0), (1, 0, 1), (1, 1, 1)\}$, and the outcomes of the remaining measurement choices maybe arbitrary for the final state (just as they are for $X \neq X'$ in Equation 13). Writing $*$ in the blocks that contain these arbitrary terms, the final state of Alice's and Bob's systems and Alice's memory after her measurement must be of the following form. Here, " stands for "this block has the same entries as the previous non-arbitrary block" and $P_{fin}(a = i, b = j, a' = k | X = l, Y = m, X' = n)$ represents the probability of obtaining the outcomes $a = i, b = j, a' = k$ when performing the fiducial measurements $X = l, Y = m, X' = n$ on the final state $\mathbf{P}_{fin}^{ABA'}$.

$$\mathbf{P}_{fin}^{ABA'} = \left(\begin{array}{c} P_{fin}(a=0, b=0, a'=0|X=0, Y=0, X'=0) \\ P_{fin}(a=0, b=0, a'=1|X=0, Y=0, X'=0) \\ P_{fin}(a=0, b=1, a'=0|X=0, Y=0, X'=0) \\ P_{fin}(a=0, b=1, a'=1|X=0, Y=0, X'=0) \\ P_{fin}(a=1, b=0, a'=0|X=0, Y=0, X'=0) \\ P_{fin}(a=1, b=0, a'=1|X=0, Y=0, X'=0) \\ P_{fin}(a=1, b=1, a'=0|X=0, Y=0, X'=0) \\ P_{fin}(a=1, b=1, a'=1|X=0, Y=0, X'=0) \\ \hline \cdot \\ \cdot \\ \cdot \end{array} \right)_{ABA'} = \left(\begin{array}{c} \frac{1}{2} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \frac{1}{2} \\ * \\ '' \\ * \\ * \\ '' \\ * \\ * \\ * \\ 0 \\ 0 \\ \frac{1}{2} \\ 0 \\ 0 \\ \frac{1}{2} \\ 0 \\ 0 \end{array} \right) \begin{array}{l} X=0, Y=0, X'=0 \\ X=0, Y=0, X'=1 \\ X=0, Y=1, X'=0 \\ X=0, Y=1, X'=1 \\ X=1, Y=0, X'=0 \\ X=1, Y=0, X'=1 \\ X=1, Y=1, X'=0 \\ X=1, Y=1, X'=1 \end{array}$$

Again, we can ignore the arbitrary blocks (marked with *) and define $\tilde{s} = i \Leftrightarrow s = s' = i$ for $s \in \{a, X\}$ and $i \in \{0, 1\}$ so that $\mathbf{P}_{fin}^{ABA'}$ is equivalent to the state $\mathbf{P}_{fin}^{\tilde{A}B}$.

$$\mathbf{P}_{fin}^{\tilde{A}B} = \left(\begin{array}{c} P_{fin}(\tilde{a}=0, b=0|\tilde{X}=0, Y=0) \\ P_{fin}(\tilde{a}=0, b=1|\tilde{X}=0, Y=0) \\ P_{fin}(\tilde{a}=1, b=0|\tilde{X}=0, Y=0) \\ P_{fin}(\tilde{a}=1, b=1|\tilde{X}=0, Y=0) \\ \hline P_{fin}(\tilde{a}=0, b=0|\tilde{X}=0, Y=1) \\ P_{fin}(\tilde{a}=0, b=1|\tilde{X}=0, Y=1) \\ P_{fin}(\tilde{a}=1, b=0|\tilde{X}=0, Y=1) \\ P_{fin}(\tilde{a}=1, b=1|\tilde{X}=0, Y=1) \\ \hline P_{fin}(\tilde{a}=0, b=0|\tilde{X}=1, Y=0) \\ P_{fin}(\tilde{a}=0, b=1|\tilde{X}=1, Y=0) \\ P_{fin}(\tilde{a}=1, b=0|\tilde{X}=1, Y=0) \\ P_{fin}(\tilde{a}=1, b=1|\tilde{X}=1, Y=0) \\ \hline P_{fin}(\tilde{a}=0, b=0|\tilde{X}=1, Y=1) \\ P_{fin}(\tilde{a}=0, b=1|\tilde{X}=1, Y=1) \\ P_{fin}(\tilde{a}=1, b=0|\tilde{X}=1, Y=1) \\ P_{fin}(\tilde{a}=1, b=1|\tilde{X}=1, Y=1) \end{array} \right)_{\tilde{A}B} = \left(\begin{array}{c} \frac{1}{2} \\ 0 \\ 0 \\ \frac{1}{2} \\ \frac{1}{2} \\ 0 \\ 0 \\ \frac{1}{2} \\ \frac{1}{2} \\ 0 \\ 0 \\ \frac{1}{2} \\ 0 \\ \frac{1}{2} \\ 0 \\ \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{array} \right)_{\tilde{A}B} \quad (16)$$

$\mathbf{P}_{fin}^{\tilde{A}B}$ is clearly the PR box distribution again but this time between Bob's system and the joint subsystem of Alice's system and memory. We can apply the exact same argument now for Bob's

measurement and we would then end up with the final state $\mathbf{P}_{fin}^{\tilde{A}\tilde{B}} = \mathbf{P}_{fin}^{AA'BB'}$ (Equation 17), which is the PR box distribution between Alice's system-memory subsystem and Bob's system-memory subsystem. Here, $\tilde{s} = i \Leftrightarrow s = s' = i$, $s \in \{a, b, X, Y\}$, $i \in \{0, 1\}$. This concludes the proof.

$$\mathbf{P}_{fin}^{\tilde{A}\tilde{B}} = \mathbf{P}_{fin}^{AA'BB'} = \begin{pmatrix} P(\tilde{a} = 0, \tilde{b} = 0 | \tilde{X} = 0, \tilde{Y} = 0) \\ P(\tilde{a} = 0, \tilde{b} = 1 | \tilde{X} = 0, \tilde{Y} = 0) \\ P(\tilde{a} = 1, \tilde{b} = 0 | \tilde{X} = 0, \tilde{Y} = 0) \\ P(\tilde{a} = 1, \tilde{b} = 1 | \tilde{X} = 0, \tilde{Y} = 0) \\ \hline P(\tilde{a} = 0, \tilde{b} = 0 | \tilde{X} = 0, \tilde{Y} = 1) \\ P(\tilde{a} = 0, \tilde{b} = 1 | \tilde{X} = 0, \tilde{Y} = 1) \\ P(\tilde{a} = 1, \tilde{b} = 0 | \tilde{X} = 0, \tilde{Y} = 1) \\ P(\tilde{a} = 1, \tilde{b} = 1 | \tilde{X} = 0, \tilde{Y} = 1) \\ \hline P(\tilde{a} = 0, \tilde{b} = 0 | \tilde{X} = 1, \tilde{Y} = 0) \\ P(\tilde{a} = 0, \tilde{b} = 1 | \tilde{X} = 1, \tilde{Y} = 0) \\ P(\tilde{a} = 1, \tilde{b} = 0 | \tilde{X} = 1, \tilde{Y} = 0) \\ P(\tilde{a} = 1, \tilde{b} = 1 | \tilde{X} = 1, \tilde{Y} = 0) \\ \hline P(\tilde{a} = 0, \tilde{b} = 0 | \tilde{X} = 1, \tilde{Y} = 1) \\ P(\tilde{a} = 0, \tilde{b} = 1 | \tilde{X} = 1, \tilde{Y} = 1) \\ P(\tilde{a} = 1, \tilde{b} = 0 | \tilde{X} = 1, \tilde{Y} = 1) \\ P(\tilde{a} = 1, \tilde{b} = 1 | \tilde{X} = 1, \tilde{Y} = 1) \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 0 \\ 0 \\ \frac{1}{2} \\ \hline \frac{1}{2} \\ 0 \\ 0 \\ \frac{1}{2} \\ \hline \frac{1}{2} \\ 0 \\ 0 \\ \frac{1}{2} \\ \hline 0 \\ \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{pmatrix} \quad (17)$$

□

D FR in measure/prepare

In PR box analysis, we encounter a peculiarity which is specific to measurement procedures in boxworld: the box “disappears” after it is measured. This can become a problem when, during the course of the experiment, the observer measuring the box has to be measured together with the box. This is the case in the original Frauchiger-Renner thought experiment. However, this issue can in principle be avoided, if one adapts the description of the experiment to the mentioned peculiarity: as soon as the agent measures the box, and it subsequently disappears, she prepares a new box for the observer on the outside to measure. For example, when Alice measures the box R , she can not only prepare a box S_a for Bob to measure (Figure 11a), but also one for Wigner, meant to contain correlations of the Bob's lab (Figure 11c). Similarly, from Bob's point of view, he prepares a box RA_b for Ursula to measure (Figure 11b); and, finally, as seen from the outside, Ursula and Wigner measure boxes RA_b and SB_a , prepared for them by Bob and Alice (Figure 11d).

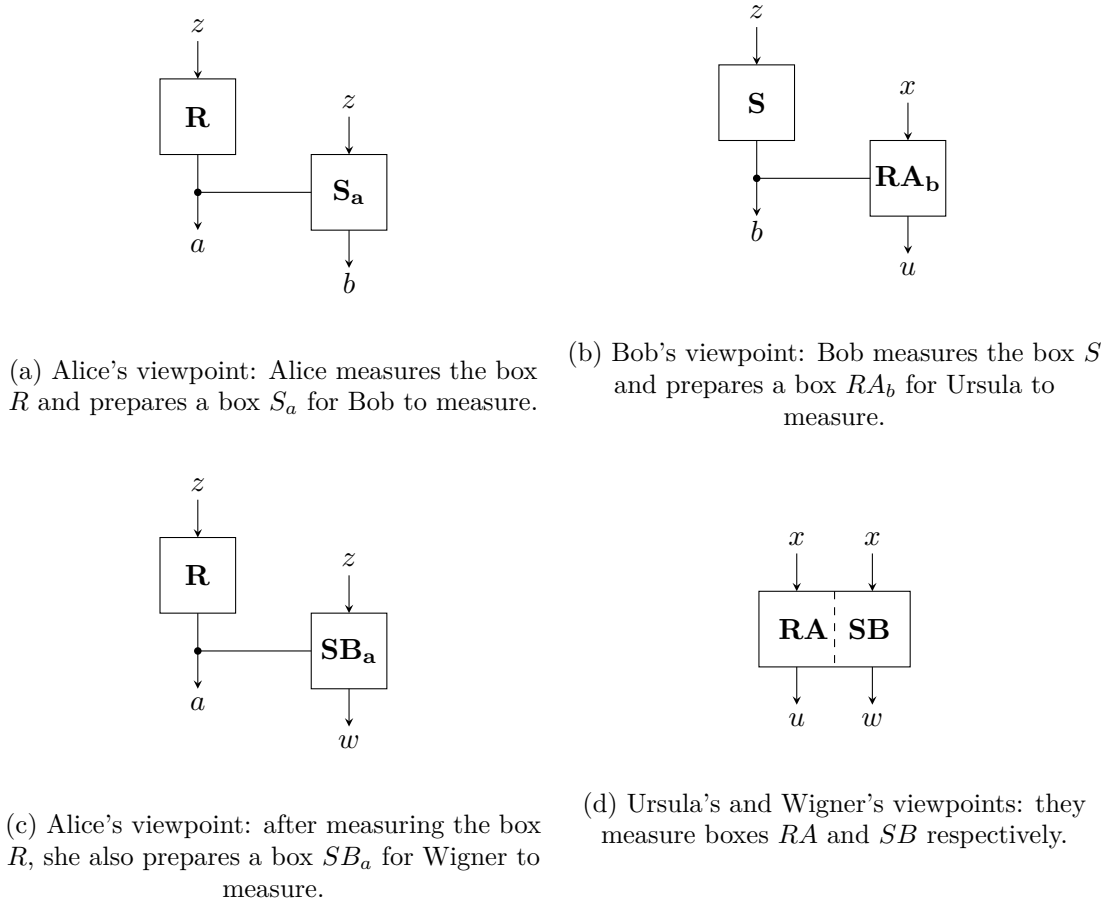


Figure 11: **Viewpoints of different agents in Frauchiger-Renner setup with boxes.**

References

- [1] Daniela Frauchiger and Renato Renner. Quantum theory cannot consistently describe the use of itself. *Nature Communications*, 9(1):3711, 2018. ISSN 2041-1723. DOI: [10.1038/s41467-018-05739-8](https://doi.org/10.1038/s41467-018-05739-8).
- [2] Lucien Hardy. Quantum theory from five reasonable axioms, 2001. [arXiv:quant-ph/0101012](https://arxiv.org/abs/quant-ph/0101012).
- [3] Jonathan Barrett. Information processing in generalized probabilistic theories. *Phys. Rev. A*, 75:032304, Mar 2007. DOI: [10.1103/PhysRevA.75.032304](https://doi.org/10.1103/PhysRevA.75.032304).
- [4] Sandu Popescu and Daniel Rohrlich. Quantum nonlocality as an axiom. *Foundations of Physics*, 24(3):379–385, Mar 1994. ISSN 1572-9516. DOI: [10.1007/BF02058098](https://doi.org/10.1007/BF02058098).
- [5] Nuriya Nurgalieva and Lidia del Rio. Inadequacy of modal logic in quantum settings. *EPCTS*, 287:267–297, 2019. DOI: [10.4204/EPTCS.287.16](https://doi.org/10.4204/EPTCS.287.16).
- [6] John Von Neumann. *Mathematical foundations of quantum mechanics*. Number 2. Princeton university press, 1955. ISBN 9780691178561.
- [7] Lucien Hardy. Nonlocality for two particles without inequalities for almost all entangled states. *Phys. Rev. Lett.*, 71:1665–1668, Sep 1993. DOI: [10.1103/PhysRevLett.71.1665](https://doi.org/10.1103/PhysRevLett.71.1665).
- [8] Samson Abramsky, Rui Soares Barbosa, Kohei Kishida, Raymond Lal, and Shane Mansfield. Contextuality, Cohomology and Paradox. In Stephan Kreutzer, editor, *24th EACSL Annual Conference on Computer Science Logic (CSL 2015)*, volume 41 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 211–228, Dagstuhl, Germany,

2015. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. ISBN 978-3-939897-90-3. DOI: [10.4230/LIPIcs.CSL.2015.211](https://doi.org/10.4230/LIPIcs.CSL.2015.211).
- [9] Robert W. Spekkens. Evidence for the epistemic view of quantum states: A toy theory. *Phys. Rev. A*, 75:032110, Mar 2007. DOI: [10.1103/PhysRevA.75.032110](https://doi.org/10.1103/PhysRevA.75.032110).
- [10] Adam Elga. Self-locating belief and the sleeping beauty problem. *Analysis*, 60(2):143–147, 2000. DOI: [10.1093/analys/60.2.143](https://doi.org/10.1093/analys/60.2.143).
- [11] Lídia Del Rio, Lea Kraemer, and Renato Renner. Resource theories of knowledge. 2015. [arXiv:1511.08818](https://arxiv.org/abs/1511.08818).
- [12] Lea Krämer and Lídia del Rio. Operational locality in global theories. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 376(2123), 2018. ISSN 1364-503X. DOI: [10.1098/rsta.2017.0321](https://doi.org/10.1098/rsta.2017.0321).
- [13] Jacob Grunhaus, Sandu Popescu, and Daniel Rohrlich. Jamming nonlocal quantum correlations. *Phys. Rev. A*, 53:3781–3784, Jun 1996. DOI: [10.1103/PhysRevA.53.3781](https://doi.org/10.1103/PhysRevA.53.3781).
- [14] Paweł Horodecki and Ravishankar Ramanathan. Relativistic causality vs. no-signaling as the limiting paradigm for correlations in physical theories, 2016. [arXiv:1611.06781](https://arxiv.org/abs/1611.06781).
- [15] David Gross, Markus Müller, Roger Colbeck, and Oscar C. O. Dahlsten. All reversible dynamics in maximally nonlocal theories are trivial. *Phys. Rev. Lett.*, 104:080402, Feb 2010. DOI: [10.1103/PhysRevLett.104.080402](https://doi.org/10.1103/PhysRevLett.104.080402).
- [16] Anthony J Short, Sandu Popescu, and Nicolas Gisin. Entanglement swapping for generalized nonlocal correlations. *Physical Review A*, 73(1):012101, 2006. DOI: [10.1103/PhysRevA.73.012101](https://doi.org/10.1103/PhysRevA.73.012101).
- [17] Paul Skrzypczyk, Nicolas Brunner, and Sandu Popescu. Emergence of quantum correlations from non-locality swapping. 2008. DOI: [10.1103/PhysRevLett.102.110402](https://doi.org/10.1103/PhysRevLett.102.110402).
- [18] Saul A. Kripke. Semantical considerations on modal logic. In *Universal Logic: An Anthology*, pages 197–208. Springer Basel, 2012. DOI: [10.1007/978-3-0346-0145-0_16](https://doi.org/10.1007/978-3-0346-0145-0_16).
- [19] James Garson. Modal logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2016 edition, 2016. URL <https://plato.stanford.edu/archives/spr2016/entries/logic-modal/>.