

General Quantum Circuit Framework for Extended Wigner’s Friend Scenarios: Logically and Causally Consistent Reasoning without Absolute Measurement Events

V. Vilasini^{1,2,*} and Mischa P. Woods^{1,2,3,†}

¹*Institute for Theoretical Physics, ETH Zurich, 8093 Zürich, Switzerland*

²*Université Grenoble Alpes, Inria, 38000 Grenoble, France*

³*ENS Lyon, Inria, LIP, 69342 Lyon, France*



(Received 23 December 2024; revised 7 October 2025; accepted 16 April 2026; published 9 September 2026)

Extended Wigner’s friend scenarios (EWFSs) model agents themselves as unitarily evolving quantum systems, going beyond the standard use of quantum theory where agents are treated classically. Several no-go results suggest deep challenges in this setting: Frauchiger and Renner argued that quantum agents reasoning using quantum theory would arrive at logical paradoxes, while the local-friendliness theorem raises difficulties for maintaining objective notions of measurement events and causal reasoning in EWFSs. This prompts a key question: Can the ability to make and test scientific predictions and reason consistently still be maintained if quantum theory applies universally, even to agents and their memories? We answer this positively by developing a general quantum circuit framework for EWFSs without assuming absolute measurement outcomes. We formalize Heisenberg cuts in terms of distinct choices of channels in the circuit, labeled by “settings,” and show that Frauchiger-Renner-type paradoxes are fully resolved by making explicit the conditioning on these channels or settings. We provide concrete reasoning rules for quantum agents that ensure logical, probabilistic, and causal consistency. Our framework captures all perspectives in an EWFS within a single causal structure while allowing events to be fundamentally subjective, and we show how objective events and predictions nevertheless emerge in current real-world quantum experiments. Thus, we demonstrate a relational yet operational framework that overcomes key challenges in EWFSs without altering the Born rule, unitarity, or classical logic and probability for outcomes. This establishes a formal platform to analyze and compare different EWFS arguments and extend quantum information methods consistently into the Wigner’s friend domain.

DOI: [10.1103/nqbv-6qgr](https://doi.org/10.1103/nqbv-6qgr)

Subject Areas: Quantum Physics

I. INTRODUCTION AND PAPER OVERVIEW

A. Introduction

Quantum theory is among the most successful physical theories, accurately describing microscopic phenomena. Experiments continue to probe quantum effects in increasingly large systems, motivated by scalable quantum computing and tests of fundamental physics. If quantum theory is universally valid, it should, at least in principle, be able to model observers or agents as physical systems within the theory.

Wigner was among the first to consider this in the 1960s through his intriguing thought experiment [1]. In it, an agent (the friend) measures a quantum system and

perceives a definite classical outcome, while an external “superagent” (Wigner) models the friend’s entire lab as a closed, unitarily evolving quantum system and could, in principle, reverse the friend’s measurement. This tension lies at the heart of the measurement problem, revealing how quantum theory applied to observers can have surprising empirical consequences (see also Ref. [2]).

More recently, there has been a surge of interest in exploring extended Wigner’s friend scenarios (EWFSs) involving more agents, some of which suggest radical implications for physics via no-go theorems. This began with an extension by Frauchiger and Renner (FR) [3], where they argued that quantum agents who also reason about each other’s knowledge using quantum theory (e.g., Born rule and unitarity) in combination with standard axioms of classical logic (e.g., distributive axiom and consistent inheritance of knowledge) applied to their observed outcomes would reach logical contradictions (see also Ref. [4]). This led to their widely cited claim that “quantum theory cannot consistently describe the use of itself.”

Other no-go results, such as Brukner’s [5,6] and the local-friendliness (LF) theorem [7], address more

*Contact author: vilasini@inria.fr

†Contact author: mischa.woods@gmail.com

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article’s title, journal citation, and DOI.

ontological aspects of quantum theory through EWFSs. The LF theorem shows that the assumptions of local agency (LA) and absoluteness of observed events (AoE) cannot be simultaneously satisfied in any theory that explains the predictions of a certain EWFS. This implies that if one wishes to preserve free choice of measurement settings and relativistic causality principles (both entailed by LA) in EWFS, then one can no longer have absolute observed events for all agents.

The nonabsoluteness of events, thus, raises concerns over whether experimental outcomes can be regarded as objective facts about the world. Meanwhile, FR's claim, taken at face value, questions the consistent usability of quantum theory by agents in EWFSs. More generally, these issues make it imperative to unambiguously and rigorously address the following question: *How can we reliably reason about the world and make consistent predictions if quantum theory applies universally to agents with full quantum control over each other's labs, without assuming absolute measurement events?*

This is relevant because a sufficiently large quantum computer could act as an agent in such EWFS arguments (see also Refs. [8,9]). EWFS no-go theorems have been met with several responses, with different interpretations of quantum theory suggesting distinct resolutions. In the FR case, responses span from conceptual clarifications [4,10,11], to new reasoning rules for avoiding paradoxes [12–14], to criticisms identifying various hidden assumptions in FR's reasoning [15–18]. Yet no general framework has emerged that can both identify which assumptions are truly necessary for the paradoxes and provide a solution to preserve logical and causally consistent reasoning.

Causality issues are central in the FR and LF scenarios, yet these have largely been analyzed independently without a clear unifying perspective. Some proposed solutions (e.g., Refs. [12,14]) to the FR paradox suggest keeping track of future operations by superagents in order to determine the validity of a measurement or its predictions made by a friend at a certain time, raising concerns: How much of the future must be tracked to reason about present outcomes, and how can agents reason about spacelike separated measurements in EWFS consistently with relativistic causality principles? The LF theorem and subsequent no-go results imply that existing quantum causal models cannot explain the predictions of such EWFSs, as they assume absoluteness of measurement outcomes [19,20]. This applies even when allowing exotic scenarios with cyclic causation, raising the challenge of whether a consistent quantum causal modeling formalism can be constructed without assuming AoE while preserving the other fundamental relativistic and operational principles such as LA.

At the core of these discussions is an ambiguity, underpinning our very understanding of what constitutes a measurement event: Should a measurement be modeled as an irreversible process yielding a classical record

(an outcome) or as a reversible unitary on a closed system? In quantum theory, this choice, linked to the Heisenberg cut, determines which parts of an experiment are treated as classical vs quantum. Wigner's scenario shows the pressing need to make this modeling explicit, which is not accounted for in existing quantum frameworks. Despite extensive work on the topic, there is still no consensus on how to consistently model EWFSs, which we argue is primarily because no prior framework exists that meets all of the following desiderata.

- (i) *Explicitly formulate predictions for EWFSs*, stating clearly the measurement modeling assumptions and knowledge conditions used and clarifying the role of Heisenberg cuts.
- (ii) *Apply to general EWFS* where agents' labs are modeled as quantum systems which can be fully quantum controlled by other agents, without assuming an absolute notion of events shared by all agents.
- (iii) *Ensure logical and causal consistency*, by providing explicit reasoning rules for such quantum agents, which are compatible with relativistic and causality principles like no-superluminal signaling.
- (iv) *Explain emergent objectivity of outcomes* in current real-world experiments and recover standard quantum predictions therein.
- (v) *Remain interpretation independent*, so that identifying relevant assumptions and resolving paradoxes does not depend on a specific interpretation of quantum theory.

The standard quantum circuit formalism satisfies such desiderata when agents are not themselves quantum systems, providing clear causal semantics, unambiguous Born-rule predictions, and interpretation-independent usage. The EWFS no-go arguments pose challenges to extending this standard formalism to Wigner's friend scenarios, which we overcome here.

B. Overview of contributions and paper layout

We extend the quantum circuit framework to EWFS enabling a relational and operational formalism for quantum measurements that answers the question posed in Sec. I A. A main contribution is that FR-type contradictions [3] do not arise in this new framework, and it, thus, resolves a challenge posed in Ref. [21]. We explicitly identify a previously implicit assumption (which we term Assumption I) responsible for the inconsistency and show that quantum mechanics remains self-consistent when this assumption is abandoned. A key ingredient of our framework, whose central importance will recur throughout the paper, is the consistent labeling of Heisenberg cuts, which we call *settings*. We now summarize our full contributions and paper layout.

- (i) *Quantum circuit framework for EWFSs*—In Sec. II, we review Wigner's thought experiment and intuitively motivate our framework through this example.

In Sec. III, we introduce our augmented quantum circuit framework for EWFSs with any number and configuration of agents, formalizing Heisenberg cuts through settings, providing a means to treat measurement events as relative rather than absolute. This enables computing consistent probabilities, demonstrating *completeness*, *consistency*, and *causality* without assuming AoE (Sec. IV A).

- (ii) *Universal resolution and identification of root cause of EWF paradoxes*—In Sec. IV B, we show that our augmented circuit enables consistent reasoning with quantum rules and classical logic. We identify Assumption I (independence of predictions from settings or Heisenberg cuts) as the failing assumption causing apparent paradoxes (Sec. IV C). Applied to the FR paradox, our framework clearly resolves it without restricting agents’ reasoning (Sec. V) and clarifies the role of Assumption I in relation to other EWFS theorems (Sec. V C).
- (iii) *Objective measurement events and causality*—Section VI distinguishes standard vs Wigner’s-friend-type quantum experiments by introducing a criterion for “superagency” based on causal structure. We show that standard scenarios recover objective measurement events independent of Heisenberg cuts.
- (iv) *Multiagent reasoning and interpretations*—Section VII A compares EWFS paradoxes with classical reasoning pitfalls, proposing a paradigm for consistent reasoning even with partial knowledge. Section VII C discusses the meaning of settings and their Bayesian-like updates, while Sec. VII B covers the interpretation independence of our results.
- (v) *Unified perspective*—Our framework provides a unified platform for multiple EWFS arguments. Appendix I relates our approach to previous responses and interpretations. In follow-up work [22], we apply our formalism to the LF scenario, linking AoE and Assumption I further.

II. MOTIVATING EXAMPLE: SIMPLEST WIGNER’S FRIEND SCENARIO

In this section, we review Wigner’s original scenario and argument [1] while highlighting the salient features and insights which are core to understanding our general framework and solution to EWFS paradoxes.

The postulates of quantum theory propose two types of evolutions: unitary evolution of closed quantum systems and the projection postulate for the evolution associated with observing a measurement outcome. However, the theory does not specify when to apply each type of evolution. While this ambiguity does not affect our ability to apply quantum theory successfully in real-world experiments (see Sec. VI for further discussion on this point),

Wigner’s 1967 thought experiment highlights beautifully why we should be concerned about this ambiguity.

The thought experiment assumes that quantum theory is universally applicable to measurement devices, agents performing the measurements, and their laboratories. Here, an agent need not be conscious or human: A sufficiently advanced quantum computer that can measure another system, store the outcome in quantum memory, and perform basic computations can serve as an agent in Wigner’s friend no-go theorems (see also Refs. [8,9,11]).

Suppose Alice (Wigner’s friend) measures a quantum system S prepared in the state $|\psi\rangle = \alpha|0\rangle_S + \beta|1\rangle_S$ in the computational basis, obtaining a classical outcome a with values $a \in 0, 1$. She stores this outcome in her memory A , initialized to the state $|0\rangle_A$. Treating A as representing the rest of Alice’s lab, SA represents Alice’s entire lab. If Alice’s lab is a closed quantum system, it evolves unitarily according to the initial premise of the universal validity of quantum theory. Thus, according to the unitarity postulate, we would describe the evolution of Alice’s lab as

$$\mathcal{M}_{\text{unitary}}^A : |\psi\rangle_S \otimes |0\rangle_A \mapsto \alpha|00\rangle_{SA} + \beta|11\rangle_{SA}, \quad (1)$$

where $|a\rangle_{SA}$ represents the effective state of Alice’s lab when she observes the outcome $a = a$. When the memory is initialized to $|0\rangle_A$, the unitary evolution $\mathcal{M}_{\text{unitary}}^A$ for the computational basis measurement on S (storing the outcome value in memory) is a CNOT gate with S as control and A as target.

Alternatively, applying the projection postulate, when Alice obtains outcome $a = 0$, her lab’s evolution is $|\psi\rangle_S|0\rangle_A \mapsto |0\rangle_S|0\rangle_A$, and, for $a = 1$, it is $|\psi\rangle_S|0\rangle_A \mapsto |1\rangle_S|1\rangle_A$. These are trace-decreasing evolutions, but summing over all possible outcomes gives the trace-preserving evolution:

$$\mathcal{M}_{\text{projection}}^A : |\psi\rangle\langle\psi|_S \otimes |0\rangle\langle 0|_A \mapsto |\alpha|^2|00\rangle\langle 00|_{SA} + |\beta|^2|11\rangle\langle 11|_{SA}. \quad (2)$$

This describes the evolution of Alice’s lab when using the projection postulate without considering specific outcomes, such as if Alice forgets the outcome after measurement.

Thus, treating a measurement as producing a classical record or as a purely unitary evolution yields two distinct trace-preserving channels for the same initial state. Wigner’s thought experiment shows that this ambiguity can have observable consequences if quantum theory is universally valid (as also emphasized in the analysis of Ref. [2]).

Consider an outside agent, Wigner, who has full quantum control over SA (Alice’s lab). Such a Wigner, a “superagent,” can perform arbitrary quantum operations on SA . Since these evolutions result in distinct states on

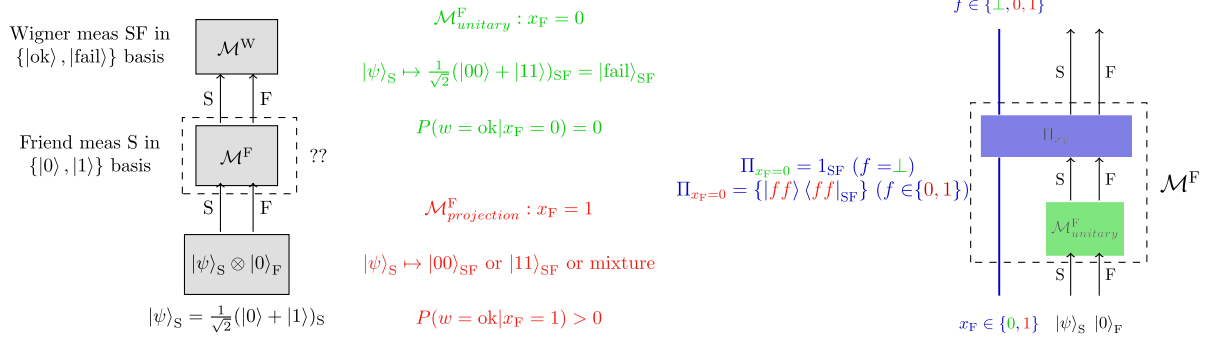


FIG. 1. Time flows from bottom to top in this figure. Left: the Friend measures a state $|\psi\rangle_S = (1/\sqrt{2})(|0\rangle + |1\rangle)_S$ in the computational basis, storing the outcome in their memory F (initially in state $|0\rangle_F$), while Wigner then measures SF in the basis $\{|ok\rangle, |fail\rangle\} = (1/\sqrt{2})(|00\rangle \mp |11\rangle)_{SF}$, yielding outcome w . If the friend’s measurement is modeled unitarily, the postmeasurement state is a pure entangled superposition (green), while if we apply the projection postulate, the postmeasurement states are different (red). Clearly, the probability of w depends on this choice (which models the Heisenberg cut), which we differentiate by the labels $x_F = 0$ and $x_F = 1$. Ignoring the conditioning on x_F would lead to an apparent contradiction: $P(w = \text{ok}) = 0$ and $P(w = \text{ok}) > 0$. Right: our framework defines an augmented circuit which incorporates the choices and resolves all EWFS paradoxes, also in more complex setups. Depicted here is an augmented version of $\mathcal{M}^F$, where a nontrivial outcome $f \in \{0, 1\}$ is possible only for $x_F = 1$ while $x_F = 0$ implies there are no nontrivial classical records ($\perp$).

SA, Wigner can operationally distinguish them by measuring SA in a suitable basis. That is, the probability of Wigner’s measurement outcome w will generally depend on how the friend Alice’s measurement is modeled: using the unitarity [Eq. (1)] or the projection postulate [Eq. (2)]. This is illustrated in Fig. 1, which also summarizes how the scenario would be modeled in our framework as an augmented circuit which incorporates this choice in how the friend’s measurement is modeled (which can be seen as a formalization of the Heisenberg cut in Sec. VIII).

Several extensions of Wigner’s original thought experiment, including Deutsch [23], FR [3], and LF [7], add further agents and/or information-processing steps. These works suggest further radical features of quantum theory exemplified by Wigner’s friend scenarios; while FR have claimed that logical contradictions arise, the LF theorem presents a case for a fundamental breakdown of absoluteness of observed events in such scenarios, despite our experience of absolute observations in existing physical experiments. In the next sections, we develop a comprehensive framework that can describe arbitrary EWFSs in quantum theory and apply it to resolve all FR-type paradoxes as well as address the emergence of absolute events.

III. A GENERAL CIRCUIT FRAMEWORK FOR EWFS

We now develop our general circuit framework for EWFS that meets all the desiderata motivated in the Introduction. We do so by carefully distinguishing between and connecting: the predictions of quantum theory, agents’ knowledge, and their statements. This is performed while defining what it means for a theoretical model to make consistent predictions in an EWFS.

A. Extended Wigner’s friend scenarios and quantum predictions

The term extended Wigner’s friend scenario is commonly used for cases such as Frauchiger-Renner and local friendliness, but a general definition has been lacking. We define EWFSs as finite multiagent quantum protocols in which agents’ memories, or, equivalently, labs, are modeled as quantum systems and one agent can have full quantum control over another’s lab. The formal definition is provided below, and its generality is justified in Appendix A.

Definition 1 (EWFS). An EWFS is a quantum protocol that consists of

- (1) a finite set $S := \{S_1, \dots, S_m\}$ of systems;
- (2) a finite set $A := \{A_1, \dots, A_N\}$ of agents;
- (3) a set $M := \{M_1, \dots, M_N\}$ of memory systems, one for each agent A_i , where they store the outcome of their measurement;
- (4) for each agent A_i , a subset $S_i \subseteq S \cup M \setminus \{M_i\}$ of systems (which can include the memories of other agents) that they measure at time t_i according to a measurement $\mathcal{M}^{A_i}$, obtaining a measurement outcome a_i that they store in their memory system M_i —moreover, $t_i < t_j$ for $i < j$;
- (5) for each agent A_i , a finite set $O_i := \{0, 1, \dots, d_{S_i} - 1\}$ in which the value a_i of their outcome a_i belongs;
- (6) for each agent A_i , a fixed operation $\mathcal{E}^i$ acting on their measured system and memory $S_i \cup \{M_i\}$ that captures the possibility of performing a further operation on the system S_i after the measurement, depending on the measurement outcome; and
- (7) a joint initial state $\rho_{S_1, \dots, S_m}$ of all the systems S at time $t_0 < t_i$ for all $i \in \{1, \dots, N\}$

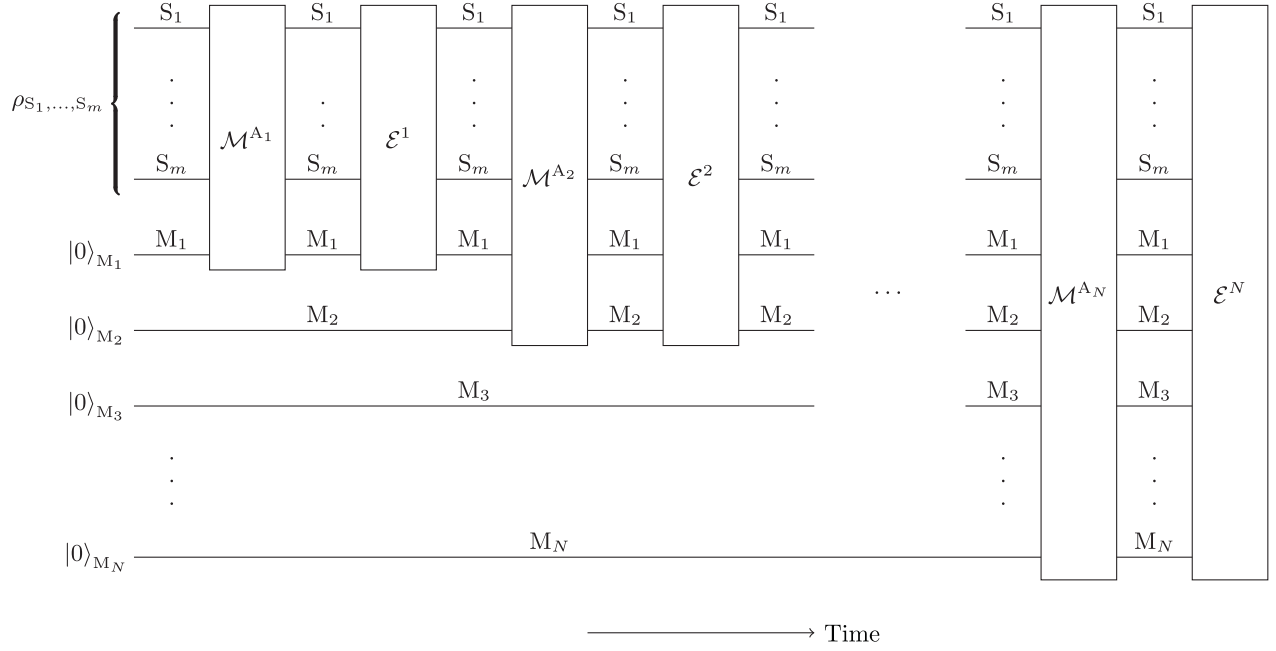


FIG. 2. General form of an EWFS involving N agents and m systems described in the main text. All agents agree on the initial state $\rho_{S_1, \dots, S_m}$ of the systems and initialize their memories to $|0\rangle$. Each agent A_i performs a measurement on some subset of the systems and on a subset of the memories of all agents who acted before, and they store the outcome of the measurement in their own memory M_i . After the measurement, each agent may perform a fixed transformation $\mathcal{E}^i$ (that is previously agreed upon by all agents) before the next agent measures. However, the agents need not necessarily agree on how each measurement is modeled; depending on their perspective, some agents may describe a measurement as a purely unitary evolution as in Eq. (3), while others may describe the same measurement as a decoherent process by assigning projectors as in Eq. (8). Because of this ambiguity in how a measurement is modeled, this diagram is not a fully specified quantum circuit.

The protocol takes the form of Fig. 2, and we can consider projective measurements $\mathcal{M}^{A_i} = \{\pi_{a_i}^{S_i} = |a_i\rangle\langle a_i|_{S_i}\}_{a_i \in \mathcal{O}_i}$ without loss of generality.

The general form of an EWFS is illustrated in Fig. 2. It is useful to note that we can assume without loss of generality that each agent performs only one measurement. Any scenario where an agent performs multiple measurements can be transformed into this form by modeling it as multiple agents, each performing a single measurement. Similarly, scenarios where an agent can choose between multiple measurements can be modeled as a single measurement by encoding the measurement choice in an initial state, as we show explicitly in an upcoming work [22] for the LF scenario, which includes different measurement choices.

In an EWFS, the unitary description $\mathcal{M}_{\text{unitary}}^{A_i}$ of a measurement $\mathcal{M}^{A_i}$ corresponds to a CNOT from S_i to the memory M_i (chosen to be of appropriate dimensions) in the basis $\{|a_i\rangle_{S_i}\}_{a_i \in \mathcal{O}_i}$ of the measurement. For any state $|\psi\rangle_{S_i} = \sum_{a_i} c_{a_i} |a_i\rangle_{S_i}$ expressed in this basis, we have [24]

$$\mathcal{M}_{\text{unitary}}^{A_i} : |\psi\rangle_{S_i} \otimes |0\rangle_{M_i} \mapsto \sum_{a_i} c_{a_i} |a_i a_i\rangle_{S_i M_i}. \quad (3)$$

As highlighted by Wigner's thought experiment, in Wigner's friend scenarios (WFSs), the ambiguity in modeling a measurement—whether through unitarity or the

projection postulate—leads to observably different predictions. Quantum theory itself does not offer a clear set of rules to resolve this ambiguity or determine the correct predictions in an EWFS. Consequently, various sets of rules can be proposed for making predictions or statements about observed outcomes in a given EWFS. Different interpretations of quantum theory yield different predictions for observable correlations in an EWFS, even though they agree on predictions for currently realisable quantum experiments.

Given this ambiguity and the potential for multiple approaches to making predictions and reasoning in an EWFS, we provide general definitions of predictions and statements. These definitions will offer a unified framework for discussing the wide range of previous results and responses to EWFS arguments and their relation to our main results. In the following, whenever we have a subset $\mathcal{M}^{A_{j_1}}, \dots, \mathcal{M}^{A_{j_p}}$ of measurements, it will be useful to denote the corresponding set of outcomes and set of values using vectors

$$\begin{aligned} \vec{a}_j &:= a_{j_1}, \dots, a_{j_p}, \\ \vec{\tilde{a}}_j &:= \tilde{a}_{j_1}, \dots, \tilde{a}_{j_p}. \end{aligned} \quad (4)$$

Then, a value assignment to a set of outcomes, such as $a_{j_1} = a_{j_1}, \dots, a_{j_p} = a_{j_p}$ becomes $\vec{a}_j = \vec{\tilde{a}}_j$.

Definition 2 (prediction and scenario parameters). Consider an EWFS along with a set $\mathcal{R}$ of rules for computing predictions in the scenario. A prediction in the EWFS is conditional probability $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, k = \mathcal{K})$, where $\vec{a}_j$ and $\vec{a}_l$ represent the set of outcomes associated with any two disjoint subsets $\mathcal{M}^{A_{j_1}}, \dots, \mathcal{M}^{A_{j_p}}$ and $\mathcal{M}^{A_{l_1}}, \dots, \mathcal{M}^{A_{l_q}}$ (latter possibly empty) of measurements in the EWFS and k is a (possibly empty) set of random variables whose values $\mathcal{K}$ encode additional information about the scenario (the exact description of which is to be specified by the rules $\mathcal{R}$). Whenever $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, k = \mathcal{K})$ takes values in $\{0, 1\}$, we refer to it as a *logical prediction*.

An example of the k parameters would be the systems, states, transformations, and measurements that one must generally condition on when computing a probability of a measurement outcome; e.g., if a state ρ_S is prepared on a system S , evolved by a transformation $\mathcal{U}$, and measured using a measurement $\mathcal{M}$ to obtain an outcome a , then the relevant outcome probability is a conditional probability $P(a | \rho_S, \mathcal{U}, \mathcal{M})$. We see more examples later.

Remark 1. In probability theory, a conditional probability is by definition “well defined.” This means it is uniquely determined by the event space and is normalized, ensuring no contradictions arise when applying them. In the above definition, however, we do not require conditional probabilities to be uniquely defined based on the rules for calculating them (e.g., if the rules are ambiguous) or to sum to one (e.g., if the rules are inconsistent). We merely assume they are numbers in $[0, 1]$, which can lead to “paradoxes” under rules that do not yield correctly normalised probabilities. We see examples where the rules produce both well-defined and not well-defined conditional probabilities.

Definition 3 (statements associated with predictions). Every prediction $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, k = \mathcal{K})$ can be associated with a corresponding statement: “If the outcomes $\vec{a}_l$ take values $\vec{a}_l$ and the additional parameters of the scenario take the value $k = \mathcal{K}$, then the outcomes $\vec{a}_j$ take values $\vec{a}_j$ with a probability $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, k = \mathcal{K})$.”

The set of all statements associated with predictions made using a set of reasoning rules in a given EWFS will be denoted as Σ (when the scenario and rules are evident from context).

Definition 4 (logical statements). Statements associated with logical predictions are called *logical statements*. When we have a logical prediction $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, k = \mathcal{K}) = 1$ (or $= 0$), the corresponding statement has the same conditional part as above, and the latter part of the statement becomes “..., then it is certain that the outcomes $\vec{a}_j$ (do not) take values $\vec{a}_j$.” For logical predictions, we can also express the statements using logical operators $\wedge$ (and), $\Rightarrow$ (implies), and $\neg$ (negation). Specifically, the statements associated with $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, k = \mathcal{K}) = 0$ and $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, k = \mathcal{K}) = 1$ would, respectively, be

$$\begin{aligned} \vec{a}_l = \vec{a}_l \wedge k = \mathcal{K} &\Rightarrow \neg(\vec{a}_j = \vec{a}_j), \\ \vec{a}_l = \vec{a}_l \wedge k = \mathcal{K} &\Rightarrow \vec{a}_j = \vec{a}_j. \end{aligned} \quad (5)$$

The set Σ_L of all such logical statements of an EWFS is a subset of Σ .

In examples, we may have $\vec{a}_l$ and k being empty sets; for instance, we can have a prediction $P(a_i = a_i) = 1$. Then, the associated logical statement is simply $a_i = a_i$ or in words “it is certain that the outcome a_i takes the value a_i .”

We then have the following definition of consistency for any set of predictive statements.

Definition 5 (consistency for Σ). A set Σ of statements obtained in an EWFS from a set of reasoning rules is said to be consistent iff Σ contains no pairs of statements S and S' associated with predictions $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, k = \mathcal{K})$ and $P'(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, k = \mathcal{K})$, respectively, where $P \neq P'$.

Lemma 1. If Σ is a set of consistent predictive statements, then

$$S \in \Sigma \Rightarrow \neg S \cap \Sigma = \emptyset, \quad (6)$$

where $\neg S$ denotes the negation of the statement S .

There are two distinct ways of being certain of an outcome value, illustrated by the following simple example.

- (i) *Scenario 1*—Alice has a fair coin yielding outcome c taking values in $\{\text{heads}, \text{tails}\}$. She flips the coin and observes $c = \text{heads}$ in a round and is then certain that $c = \text{heads}$.
- (ii) *Scenario 2*—Alice has a biased coin that reads heads on both sides, $P(c = \text{heads}) = 1$, and she is certain that $c = \text{heads}$ in every round (without observing the outcome).

Clearly, these scenarios are different. To ensure reliability and consistency when communicating with other agents, it is necessary to distinguish these cases. Our previous definition of statements associated with predictions covers only scenario 2. To cover scenario 1, we need to include a set of statements associated with observations that agents make in a given experimental run, denoted as Σ_{obs} . We define this case in Appendix B and also discuss a more refined definition of consistency that accounts for both predictive and observational statements there. For all the main results and discussions, it will be sufficient to consider the predictive statements Σ and the generalization of the results to include Σ_{obs} is presented in Appendix B. This is because a theory permits observation of certain outcomes only when its predictions assign a nonzero probability to that outcome; thus, the main features of the theory can be understood by studying its predictions and associated statements as defined here.

B. Formulating an EWFS within a single quantum circuit

As stressed before, in EWFS, there is ambiguity in applying the postulates of quantum theory to compute probabilities. The postulates can be applied in different ways (using chosen “rules”) to obtain different probabilities for the same outcomes, while distinct rules could still lead to the same probabilities in certain cases. In previous literature on EWFS, there is a common pattern in the probabilities used in EWFS arguments, such as in the FR and LF results. However, the rules for arriving at these probabilities are often not explicitly specified.

We consider the probabilities conventionally used in previous literature when they refer to “predictions of quantum theory” in the context of EWFS, where it is assumed that unitary evolutions can be applied to agents’ labs [25]. These probabilities typically do not condition on additional variables k or channels in the circuit; their expressions refer to only the outcomes—e.g., $P(a=0, b=1)$ denotes the probability of the measurement outcomes $a=0$ and $b=1$ (which can be associated with agents Alice and Bob) in a given scenario. Below, we formalize one way to arrive at these type of probability expressions in an EWFS, which recovers the probabilities used in the FR and LF scenarios.

Definition 6 (conventional predictions in an EWFS). A conventional prediction in an EWFS is a conditional probability $P_{\text{conv}}(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l)$ evaluated by applying the following rules.

- (1) In the given EWFS (see also Fig. 2), all the measurements $\mathcal{M}^{A_i} \notin \{\mathcal{M}^{A_{j_1}}, \dots, \mathcal{M}^{A_{j_p}}\} \cup \{\mathcal{M}^{A_{l_1}}, \dots, \mathcal{M}^{A_{l_q}}\}$ (those not appearing in the prediction) are modeled as unitary evolutions of their labs [according to the channel Eq. (3)].
- (2) For each $\mathcal{M}^{A_i} \in \{\mathcal{M}^{A_{j_1}}, \dots, \mathcal{M}^{A_{j_p}}\} \cup \{\mathcal{M}^{A_{l_1}}, \dots, \mathcal{M}^{A_{l_q}}\}$, the projective measurement $\{\pi_{a_i}^{S_i} = |a_i\rangle\langle a_i|_{S_i}\}_{a_i \in \mathcal{O}_i}$ is applied on the corresponding system S_i , followed by a CNOT in the same basis from the system S_i to the memory M_i which models the procedure of making the measurement on the system and storing the outcome in the memory.
- (3) This fixes all the channels in the circuit, and the joint probability $P_{\text{conv}}(\vec{a}_j = \vec{a}_j, \vec{a}_l = \vec{a}_l)$ is then calculated by applying the Born rule to this circuit, using the measurement projectors given above [26].
- (4) The prediction $P_{\text{conv}}(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l)$ is then obtained through the usual conditional probability rule.

A conventional prediction corresponds to the case where k is the empty set, meaning the prediction is not conditioned on any additional information about the scenario (such as states and channels used). Such predictions often lead to FR-type paradoxes, highlighting that the above rules for computing probabilities in EWFS are not generally

consistent and do not yield valid joint probabilities for measurement outcomes in EWFSs.

In standard quantum theory outside the context of WFS, where agents’ labs are not typically modeled as quantum systems in a circuit, all predictions for a given experiment are defined relative to a single circuit, ensuring consistent and well-defined joint probabilities. In EWFS, it is initially unclear which circuit is associated with a given scenario due to the ambiguity in the channels modeling the measurements (as discussed in Sec. II), which has observable consequences in these scenarios. We now explain how every EWFS can be mapped to a single quantum circuit which we call the *augmented circuit* of the EWFS, from which all the predictions of the EWFS can be computed and used consistently, while allowing measurements to be modeled as unitary evolutions of agents’ labs and also preserving the validity of the Born rule.

For this, recall that for a measurement $\mathcal{M}^{A_i}$ in a basis defined by the projectors $\{\pi_{a_i}^{S_i} = |a_i\rangle\langle a_i|_{S_i}\}_{a_i \in \mathcal{O}_i}$ acting on a state $|\psi\rangle_{S_i} = \sum_{a_i} c_{a_i} |a_i\rangle_{S_i}$, the associated unitary evolution $\mathcal{M}_{\text{unitary}}^{A_i}$ is as given in Eq. (3). On the other hand, applying the projection postulate for the measurement on S_i and the fact that the outcome is copied to the memory A_i , one would obtain the final state $|a_i a_i\rangle\langle a_i a_i|_{S_i M_i}$ when the outcome is a_i . If we consider the view that such a classical outcome is obtained but one lacks knowledge of its value, we would obtain the following trace-preserving evolution:

$$\begin{aligned} \mathcal{M}_{\text{projection}}^{A_i}(|\psi\rangle\langle\psi|_{S_i} \otimes |0\rangle\langle 0|_{M_i}) \\ = \sum_{a_i} |c_{a_i}|^2 |a_i a_i\rangle\langle a_i a_i|_{S_i M_i}. \end{aligned} \quad (7)$$

Notice that $\mathcal{M}_{\text{projection}}^{A_i}(|\psi\rangle\langle\psi|_{S_i} \otimes |0\rangle\langle 0|_{M_i})$ above can be equivalently expressed as

$$\sum_{a_i} \pi_{a_i}^{S_i M_i} \left(\mathcal{M}_{\text{unitary}}^{A_i}(|\psi\rangle\langle\psi|_{S_i} \otimes |0\rangle\langle 0|_{M_i}) \mathcal{M}_{\text{unitary}}^{A_i \dagger} \right) \pi_{a_i}^{S_i M_i}, \quad (8)$$

where $\{\pi_{a_i}^{S_i M_i} := |a_i a_i\rangle\langle a_i a_i|_{S_i M_i}\}_{a_i \in \mathcal{O}_i}$ and we still have a one-to-one correspondence between the measurement outcomes a_i and an element from the above set of projectors. Therefore, the two possible (trace-preserving) evolutions associated with a measurement $\mathcal{M}^{A_i}$ are now given by Eqs. (3) and (8). They differ is whether or not one views the measurement as having produced classical records.

We explicitly account for this choice by introducing a classical binary variable for each measurement that takes values $x_i \in \{0, 1\}$ (which we call the setting), such that $x_i = 0$ and $x_i = 1$ correspond to the evolution equations (3) and (8), respectively.

This is captured through the following enlarged set of projectors, obtained by appending a trivial outcome value $\perp$ to the outcome set $\mathcal{O}_i$:

$$\Pi_{x_i}^{A_i} := \begin{cases} \{\pi_{a_i, x_i}^{A_i}\}_{a_i \in \{\perp\}} = \{\pi_{\perp, 0}^{A_i} := 1_{S_i M_i}\} & \text{if } x_i = 0, \\ \{\pi_{a_i, x_i}^{A_i}\}_{a_i \in \mathcal{O}_i} = \{\pi_{a_i, 1}^{A_i} := \pi_{a_i}^{S_i M_i}\}_{a_i \in \mathcal{O}_i} & \text{if } x_i = 1. \end{cases} \quad (9)$$

That is, for $x_i = 0$, the value of the outcome a_i is fixed uniquely to $a_i = \perp$, and the corresponding projector is the identity operator $1_{S_i M_i}$. Meanwhile, for $x_i = 1$, a_i takes values a_i in the set $\mathcal{O}_i$ (reflecting the different possible classical outcomes of the measurement; see Definition 1). We use the terminology *outcome* to refer to only the value of a_i when $x_i = 1$. In the case $x_i = 0$, $a_i = \perp$ and we do not regard this as an outcome (we sometimes refer to this case as the *trivial outcome*).

Then, we can model each measurement $\mathcal{M}^{A_i}$ as the corresponding unitary $\mathcal{M}_{\text{unitary}}^{A_i}$, followed by a (setting-dependent) projective measurement $\Pi_{x_i}^{A_i}$. The pure unitary picture is always recovered by setting $x_i = 0$, in which case only $\mathcal{M}_{\text{unitary}}^{A_i}$ unitary is applied. The set of choices of

settings for all N agents can be represented by a vector $\vec{x} = (x_1, \dots, x_N)$. With this explicit model, we can transform any EWFS into a standard temporally ordered quantum circuit parametrized by the settings $\vec{x}$ such that all the predictions made by different agents having different perspectives can be derived from the single circuit by fixing the settings $\vec{x}$ (as shown in Theorem C1). We call this the *augmented circuit* of the EWFS. This is illustrated in Fig. 3 and summarized in the following definition.

Definition 7 (augmented circuit of an EWFS). Given an EWFS of the form of Fig. 2, we can associate with it an augmented circuit of the form of Fig. 3 which is obtained from the EWFS by replacing each measurement $\mathcal{M}^{A_i}$ by the corresponding unitary description $\mathcal{M}_{\text{unitary}}^{A_i}$ [Eq. (3)], followed by the setting-dependent enlarged set of projectors $\Pi_{x_i}^{A_i}$ [Eq. (9)]. The setting takes binary values $x_i \in \{0, 1\}$, where we have only the trivial outcome $\perp$ whenever $x_i = 0$ and the nontrivial measurement outcome $a_i \in \mathcal{O}_i$ whenever $x_i = 1$. We refer to this as an augmented EWFS, in short.

We now make a crucial observation. While we have so far given full freedom in choosing how to model the measurements, with both $x_i = 0$ (pure unitary evolution) and $x_i = 1$ (also assigning nontrivial projectors) being

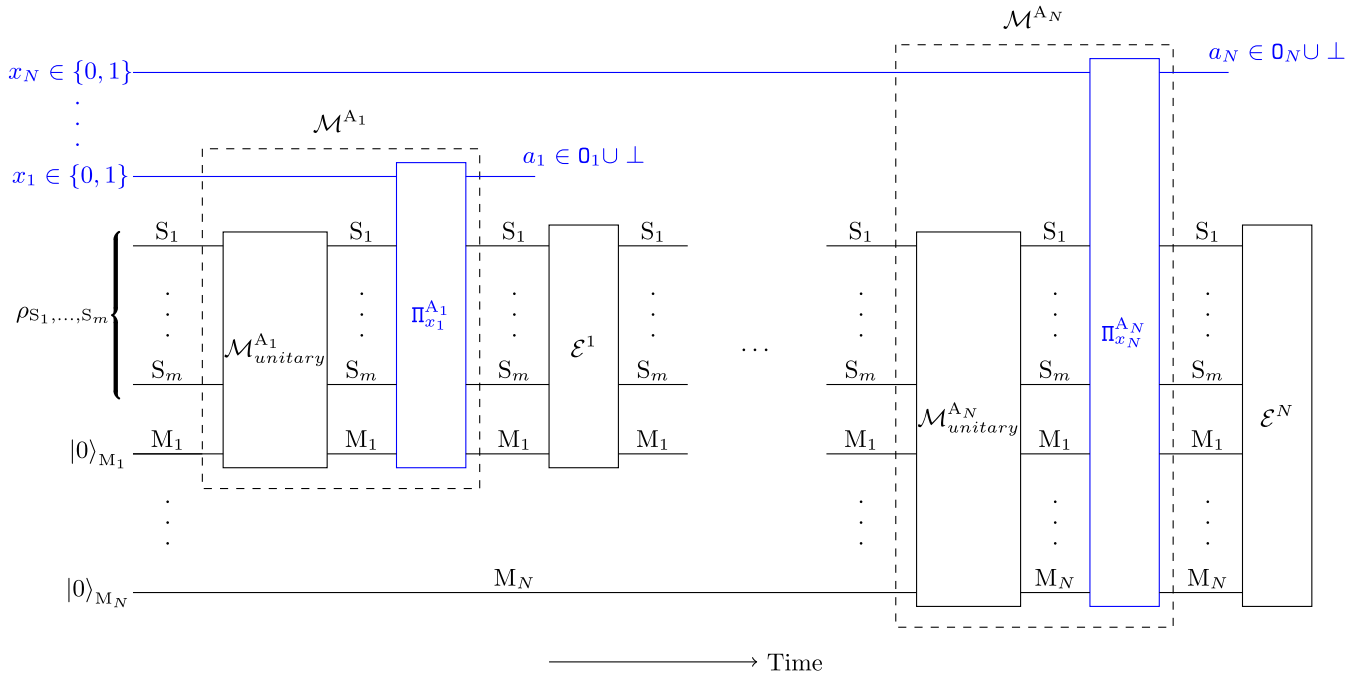


FIG. 3. Augmented circuit for the general form of an EWFS illustrated in Fig. 2 that makes explicit the implicit setting choices needed to model the measurement of each agent. This circuit makes it clear that each agent has a choice in how they describe each measurement in the scenario; when setting $x_i = 0$, the corresponding measurement $\mathcal{M}^{A_i}$ is modeled as a unitary evolution $\mathcal{M}_{\text{unitary}}^{A_i}$ (in this case, the projector in the blue box implements an identity operation and $a_i = \perp$ deterministically), and when $x_i = 1$, the same measurement is modeled as the unitary $\mathcal{M}_{\text{unitary}}^{A_i}$ followed by nontrivial projectors that identify the classical measurement outcome $a_i \in \{0, 1, \dots, d_{S_i} - 1\} := \mathcal{O}_i$. In order for any agent to reason about the measurement outcome a_i of an agent A_i , they must necessarily choose $x_i = 1$ in order to identify and calculate probabilities for the classical outcome that they are reasoning about. They may, however, choose to model all other agents A_j unitarily by choosing $x_j = 0$.

allowed, we notice that the form of the prediction being computed using the Born rule fixes some of these choices. For example, consider a prediction $P(a_i = a_i | \vec{x})$. To compute this using the Born rule, we must apply the projector associated with the outcome $a_i = a_i$, which is needed to identify that outcome. Therefore, for this prediction, $x_i = 1$. The other components of the setting vector $\vec{x}$ can generally vary according to the rules of reasoning being applied.

When we say “applying a projector,” this is not to be conflated with “collapsing” the state. Even when we do not explicitly refer to the postmeasurement state of an agents’ measurement, when reasoning about that agents’ outcome using the Born rule, we apply knowledge of the measurement basis (given here by the projectors which identify the outcome). This is further discussed in Remark 3, where we show how our framework can also be applied to EWFSs where the Born rule is applied without need for a particular state update rule. The main observation above is that, given a prediction, the measurements whose outcomes appear in that prediction are always regarded as producing classical records (associated with $x_i = 1$ in our framework) when computing the prediction.

Applying this simple observation about quantum probabilities to agents’ reasoning, we can consider an example. If A_i reasons about A_j ’s outcome based on their own outcome, they set $x_i = x_j = 1$ in order to identify the classical outcomes they are reasoning about, but they may choose $x_k = 0$ for all $k \neq i, j$ and model all other agents unitarily. How these remaining settings are to be chosen will need to be specified by a set of reasoning rules $\mathcal{R}$, which we do not yet fix. This generality will become relevant when discussing the different responses to FR’s no-go theorem (Appendix I).

Having defined the augmented circuit, we explain how one can compute predictions here. This can be done by applying the Born rule along with the conditional probability rule, as long as we are given a choice of initial values for the setting vector $\vec{x}$ (as this is required for the full specification of the circuit).

Different interpretations of quantum theory would generally propose different rules $\mathcal{R}$ to fully specify $\vec{x}$ and can arrive at different predictions for the same EWFS. As we see in Theorem C1, formalizing conventional predictions (Definition 6) in our augmented circuit yields an explicit rule for choosing the settings which will be relative to a subset of systems of measurements being considered and can, therefore, allow subjective choices of settings when we consider reasoning agents (see also Sec. VII A).

Until then, all our results regarding the augmented circuit apply to all possible rules for choosing the settings and, hence, apply to several interpretations of quantum theory (both relational and nonrelational ones).

Definition 8 (setting-conditioned prediction). Given an EWFS, a setting-conditioned prediction is a prediction

$P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi})$ (Definition 2), where the setting vector $\vec{x}$ represents the set k of scenario parameters. A setting-conditioned prediction is computed by mapping the EWFS to an augmented circuit and then applying the quantum Born rule together with the rule for conditional probabilities for the given choice $\vec{x} = \vec{\xi}$ of the settings. This is explicitly shown in Appendix E. Whenever $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi}) \in \{0, 1\}$, we refer to it as a *logical setting-conditioned prediction*.

Definition 9 (statements associated with setting-conditioned predictions). Every setting-conditioned prediction $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi})$ can be associated with a corresponding statement, in the same way that Definition 3 assigns statements to predictions $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, k = \ell)$, but for setting-conditioned predictions, the scenario parameter values $k = \ell$ are replaced with the setting values $\vec{x} = \vec{\xi}$. Statements associated with logical setting-conditioned prediction can be expressed using logical operators as given in Definition 4 but with $k = \ell$ replaced by $\vec{x} = \vec{\xi}$.

Definition 10 (set of all statements in an augmented EWFS). Consider the set of all setting-conditioned predictions that can be made in an augmented circuit of an EWFS, obtained under all possible setting choices. The set of all statements Σ^{aug} in that EWFS is obtained by mapping each setting-conditioned prediction to a corresponding statement as per Definition 9. Similarly, restricting to the set of all logical predictions, we have the corresponding set of all logical statements of the scenario which are denoted as $\Sigma_L^{\text{aug}} \subseteq \Sigma^{\text{aug}}$.

Remark 2. In setting-conditioned predictions, the setting variables are fixed to some specific value deterministically. More generally, we can consider arbitrary prior distribution over the settings $\vec{x}$ in our framework. This allows one to compute a probability distribution such as $P(\vec{a}_j = \vec{a}_j)$ which does not explicitly feature any settings, which is obtained by choosing some prior $P(\vec{x})$ and averaging over the settings, $P(\vec{a}_j = \vec{a}_j) = \sum_{\vec{\xi}} P(\vec{a}_j = \vec{a}_j | \vec{x} = \vec{\xi}) \times P(\vec{x} = \vec{\xi})$.

We have established that, for statements made using setting-conditioned predictions, the specific setting value assumed must be explicitly specified. Analogously, our framework and results can be generalized to accommodate arbitrary priors by ensuring that, when using such priors to compute outcome probabilities or predictions, the corresponding statements about the measurement outcomes specify the chosen prior over the settings, as this is an additional choice that an agent must make in the reasoning process. However, we do not delve into the case of general prior distributions further, as it is not pertinent to the main results and insights of our work.

IV. COMPLETENESS, CONSISTENCY, AND CAUSALITY WITHOUT ABSOLUTE EVENTS

In this section (and in Appendix C), we formalize and prove several key properties of our framework that are pertinent to logical and causal reasoning in EWFS, without imposing an absolute notion of measurement events.

A. Properties of general quantum predictions in EWFS

We begin by emphasizing that our framework does not assume an absolute, objective notion of measurement events, unlike most existing quantum information frameworks. This is because it does not require a global joint probability distribution $P(a_1, \dots, a_N)$ over the nontrivial outcomes $a_i \in \mathcal{O}_i$ of all measurements. Instead, by introducing settings $\vec{x}$, which as discussed in Sec. VII C, model choices of Heisenberg cuts, the framework allows outcome probabilities to be fundamentally relational in Wigner's friend scenarios.

Consider an EWFS with two measurements $\mathcal{M}^{A_1}$ and $\mathcal{M}^{A_2}$. If the joint system $S_1 M_1$ (modeling agent A_1 's lab) is treated as recording classical outcomes, while $S_2 M_2$ evolves unitarily, then $x_1 = 1$ and $x_2 = 0$, and we can compute probabilities only for $a_1 \in \mathcal{O}_1$. If both measurements are treated as classical, $x_1 = x_2 = 1$, enabling computation of joint probabilities for a_1 and a_2 . These probabilities may differ, as different x_2 values are used. Nonetheless, both are derived from the same quantum circuit via different choices of the input variable x_2 .

Although our formalism does not impose event absoluteness, its formulation as a well-defined quantum circuit (Sec. III) immediately implies that several key properties underlying scientific reasoning and important physical principles remain valid in Wigner's friend scenarios. Full definitions and theorems are provided in Appendix C 1, while the main ideas are outlined here at a more intuitive level.

In Appendix C 1 and in Theorem C1, we formally define and prove that our augmented circuit framework respects the following three properties, implying that these properties can be rigorously ensured in any EWFS in quantum theory.

- (i) *Completeness*—For any given EWFS, the corresponding augmented circuit uniquely specified by Sec. III encodes all the conventional predictions of that EWFS: For each conventional prediction, there exists a corresponding setting-conditioned prediction in the augmented circuit that is equal to it.
- (ii) *Consistency*—For any EWFS, the set of all statements Σ^{aug} obtained in the corresponding augmented circuit (Definition 10) are consistent according to Definition 5.
- (iii) *Causality*—For any EWFS, the corresponding augmented circuit is such that the associated probabilities of any set of outcomes does not depend on

setting choices made outside its past. Here, the past is defined relative to the connectivity of the circuit which is consistent with the temporal order: One operation causally precedes another when there is a path in the circuit connecting an output of the first to an input the second, and this is possible only when the second appears later in time than the first (Definition C1).

These properties confirm that our framework preserves consistency in any EWFS without violating basic physical principles like causality (neither in the quantum information nor the relativistic sense of the word).

B. Application to agents' reasoning

So far, our results have been about the properties of predictions (probabilities) that can be computed in an EWFS and showing that our formalism yields a complete, logical, and causally consistent way to make predictive statements in such scenarios. We now apply our general formalism to reveal properties of agents' reasoning that hold in general EWFS. One immediate corollary of our general consistency result of Theorem C1 for agents' reasoning is the following.

Corollary 1. If any two agents use the same choice of settings $\vec{x}$ for all measurements $\{\mathcal{M}^{A_i}\}_i$ in an augmented EWFS, then they make all the same predictions in that scenario.

This corollary concerns a rather restricted case where all agents have a fixed and common choice of Heisenberg cut (here, formalized through the settings). More generally, our framework allows agents to choose different settings depending on their perspective or prediction they wish to compute. Below, we establish that our formalism allows quantum agents in an EWFS to reason consistently even when they apply typical axioms of classical logic to observed classical outcomes, the Born rule to compute probabilities, can model each others' labs as unitarily evolving quantum systems and have fundamentally subjective perspectives. In particular, we provide a formalization of the assumptions Q , U , C , D , and S used in FR's arguments [3,4] which FR have claimed to be inconsistent in a particular EWFS [27]. Contrary to FR's conclusions, by formalizing these assumptions in our framework as $\mathbf{Q}$, $\mathbf{U}$, $\mathbf{C}$, $\mathbf{D}$, and $\mathbf{S}$, we prove that all of them can be consistently satisfied in any EWFS. The formal definitions are provided in Appendix C 2; we summarize the intuition and state the main results here.

- (i) $\mathbf{Q}$ —Agents involved in any EWFS take statements of the form $S :=$ "If the outcomes $\vec{a}_l$ take values $\vec{a}_l$ and the settings take the value $\vec{x} = \xi$, then the outcomes $\vec{a}_j$ take values $\vec{a}_j$ with a probability P " to be true when the probability that $\vec{a}_j = \vec{a}_j$ conditioned on $\vec{x} = \xi$ and $\vec{a}_l = \vec{a}_l$ computed using the quantum mechanical Born rule in that EWFS equals P .

- (ii) **U**—An agent A_i can model the measurement of another agent A_j as a unitary evolution of A_j 's lab (i.e., A_i can choose the setting $x_j = 0$ in certain predictions they make).
- (iii) **C**—For any given EWFS, and any statement $S \in \Sigma^{\text{aug}}$ possible therein, if an agent A_j knows that an agent A_i knows the statement S , then A_j can conclude that they know the statement S . In other words, agents can freely inherit each others' knowledge of statements derived from the augmented circuit in any EWFS.
- (iv) **D**—For any given EWFS, and any logical statements $S_1, S_2 \in \Sigma_L^{\text{aug}}$ possible therein, if an agent A_i knows that S_1 is true and that S_1 implies S_2 is true, then they can conclude that S_2 is true.
- (v) **S**—If a subset $\vec{a}_j$ of measurement outcomes in an EWFS is known with certainty to take the values $\vec{a}_j$, then it cannot be known with certainty that $\vec{a}_j$ takes any other values $\vec{a}_j' \neq \vec{a}_j$.

We then obtain the following corollary, which follows by construction of our framework along with the general consistency result of Theorem C1. We nevertheless include a proof in Appendix J for completeness.

Corollary 2. If agents in an EWFS reason about each other's knowledge using the augmented circuit for the scenario, then they can never arrive at a logical contradiction even if they reason using all five assumptions **Q**, **U**, **C**, **D**, and **S**.

We have given a formalization **Q**, **U**, **C**, **D**, and **S** of the FR assumptions **Q**, **U**, **C**, **D**, and **S** within our framework, showing our version of the five assumptions to be perfectly consistent in an EWFS, despite FR's claim that the original version of these assumptions lead to contradictions. FR's assumptions, although motivated as capturing “quantum theory” and “logical axioms,” do not appear to be fully and rigorously formalized, allowing room for interpretation (see also Appendix I for references to previous responses to FR's arguments). See Sec. V C for further discussion on the interpretation of FR's claims in light of our results.

Physically, our assumptions **Q**, **U**, **C**, **D**, and **S** still encompass the same essential physical requirements highlighted by FR: the universal applicability of quantum theory (Born rule + unitarity), the inheritance of agents' knowledge, and the validity of classical logic applied to knowledge of measurement outcomes. However, mathematically, within the Kripke structure of modal logic, the set of statements Σ to which our assumptions apply differs from that considered in the original modal logic formulation of FR's result (as given in Ref. [4]), due to the additional structure provided by the setting labels in our framework.

In the forthcoming section, we identify an additional assumption within our framework necessary to reproduce apparent inconsistencies akin FR's. This underscores that an implicit assumption concerning the independence of

predictions from the choice of Heisenberg cuts is required to establish an FR-type no-go theorem in quantum theory.

C. Reason for apparent inconsistencies

Despite satisfying a formal version of each of the FR assumptions (Theorem C1 and Corollary 2), our framework remains logically consistent. This raises the question: What additional assumption is needed to reproduce logical contradictions such as the apparent FR paradox within our framework?

The main difference between our framework and previous analyses of EWFSs is in the explicit introduction of the settings. Here, we say “explicit” since the settings are indeed present in the conventional computations of quantum predictions, as a choice about how measurements are modeled has to be made when computing the probabilities. The difference between these conventional predictions and statements and ours is that the former do not specify this in the probability expressions or at the level of statements being made, while our formalism does so. Dropping this choice corresponds to an assumption regarding the ability to ignore the setting choice or the Heisenberg cut. We formalize this assumption below within our framework and show that this is necessary to recover a paradox. This, in turn, yields a more precise and refined interpretation of FR-type apparent paradoxes (as discussed in Sec. V C after analyzing the example of the FR scenario). We also discuss its relation to other results in the literature and interpretations of quantum theory in Appendix I.

Definition 11 (Assumption I: Independence). A setting-conditioned prediction $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi})$ is said to be setting independent if $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi}) = P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi}')$ for all allowed values ξ and ξ' of the settings [29]. Then, the prediction can be consistently represented by dropping the settings, as $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi}) = P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l)$.

We can then immediately obtain the following corollary of our results.

Corollary 3. In order to obtain an apparent logical contradiction (i.e., a violation of **S**) in an augmented EWFS where agents reason using assumptions **Q**, **U**, **C**, and **D**, it is necessary to assume **I** on at least one logical setting-conditioned prediction that is not setting independent.

This corollary follows because first, using Corollary 2, we have that the augmented EWFS is perfectly consistent with all five assumptions **Q**, **U**, **C**, **D**, and **S**. Furthermore, the consistency of our framework captured by Theorem C1 and Corollary 1 implies that any apparent violation of **S** [i.e., $P(\vec{a}_j = \vec{a}_j) = 1$ and $P(\vec{a}_j = \vec{a}_j') = 1$ for $\vec{a}_j' \neq \vec{a}_j$] that might be obtained in an EWFS (for instance, the violation obtained by FR [3]) necessarily arises by computing the probability of the outcomes under two distinct choices of settings and then ignoring this choice by

identifying the two predictions as the same (which is equivalent to applying **I** to the two setting-conditioned predictions).

In other words, an apparent violation of **S** such as $P(\vec{a}_j = \vec{a}_j) = 1$ and $P(\vec{a}_j = \vec{a}'_j) = 1$ when formulated explicitly within our framework always translates to $P(\vec{a}_j = \vec{a}_j | \vec{x} = \vec{\xi}) = 1$ and $P(\vec{a}_j = \vec{a}'_j | \vec{x} = \vec{\xi}') = 1$ (where $\vec{\xi}$ and $\vec{\xi}'$ are two distinct setting values), which is not paradoxical as it refers to two distinct conditional probability distributions. Note, however, that this necessarily violates **I**, since the predictions are indeed dependent on the setting, which is why the apparent paradox is recovered when **I** is imposed: The contradiction arises from imposing **I** in a scenario where it fails to hold.

V. A SIMPLE RESOLUTION TO THE FR APPARENT PARADOX

Having developed a fully general framework, we now apply it to a concrete example, to demonstrate a simple resolution of the original FR paradox. We focus on the entanglement version of the FR protocol, illustrated in Fig. 4 and reviewed in detail in Appendix D 2, highlighting the core idea behind the resolution at both the conceptual and formal level. This version, proposed in talks by

Masanes and Pusey, is analyzed in full detail in Appendix G 1, which includes all supporting calculations.

While many view the entanglement version as equivalent to the original prepare-and-measure protocol, the authors of the FR paper have emphasized that it omits key subtleties involving timing, of the measurements and reasoning steps, which were central to the original formulation. Our framework is fully general and resolves the apparent paradoxes in both versions in the same way. In Appendix G 2, we address the original FR paradox directly, providing a step-by-step analysis showing how the paradox entirely disappears within our framework. This holds even when all agents reason freely using unitary quantum theory, the standard Born rule, and classical logic, and all statements from FR's original argument are preserved in a more explicit form, in our framework.

Key takeaway—FR's paper argued that “quantum theory cannot consistently justify the use of itself,” based on a paradox involving quantum agents reasoning about others' measurement outcomes. Our resolution reveals a key hidden assumption in this argument: assumption **I** (setting independence). The contradiction arises only when quantum theory is used in a restricted way that imposes **I** [30]. Observing that the actual protocol violates this assumption—i.e., predictions do depend on settings (which describe whether a measurement is modeled purely

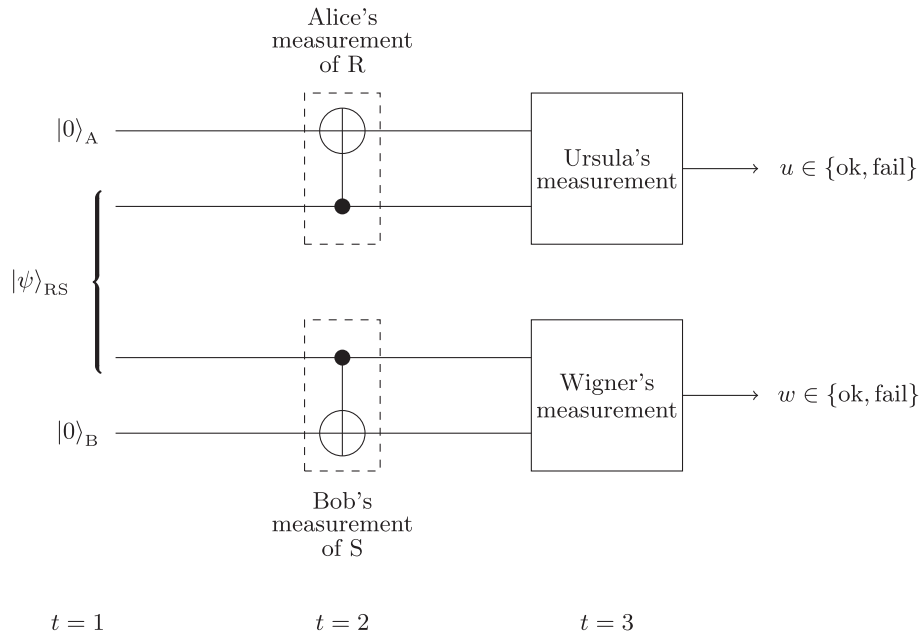


FIG. 4. Circuit that describes the entanglement version of the FR protocol, from the view of the superagents Ursula and Wigner who describe the measurement of Alice and Bob as unitary evolutions. The protocol proceeds as follows: Alice and Bob share a bipartite state $|\psi\rangle_{RS} = (1/\sqrt{3})(|00\rangle + |10\rangle + |11\rangle)_{RS}$. Alice measures R and Bob measures S, both in the computational basis, and the agents store the outcome of the measurement in their memories A and B, respectively (the unitary description of these measurements is a CNOT). Ursula then measures RA (Alice's lab) and Wigner measures SB (Bob's lab); both agents measure in the $\{|\text{ok}/\text{fail}\rangle := (1/\sqrt{2})(|00\rangle \mp |11\rangle)\}$ basis to obtain the outcomes $u, w \in \{\text{ok}, \text{fail}\}$. Note that this circuit alone does not allow the superagents to reason about the classical measurement outcome of Alice and Bob, as these are modeled purely unitarily, and, hence, no classical measurement outcomes a and b are identified.

unitarily or as producing classical records)—we relax **I** and show that the paradox vanishes. By tracking the settings in the augmented circuit associated with each agent’s reasoning, we provide explicit versions of their statements, which are consistent with quantum theory and classical logic, but their combination no longer yields any contradictions.

A. Augmented circuit

We first formulate the entanglement version of the FR protocol within our framework by giving its augmented circuit. A typical circuit associated with this scenario, also found in the previous literature, is shown in Fig. 4. The caption of the figure gives a quick recap of the protocol sufficient to follow this discussion. It is one which models the measurements of the agents Alice and Bob purely unitarily. As these are computational basis measurements, the corresponding unitary is the CNOT. However, this circuit alone does not allow us to calculate the probabilities of Alice and Bob’s classical outcomes. In the previous literature (see, for instance, Ref. [11]), multiple different circuits are considered for calculating the probabilities from the perspective of each agent. In our framework, we have shown that all such reasoning can, in fact, be captured with a single circuit—the augmented circuit. This circuit includes a setting x_i for each agent A_i , which when set to $x_i = 0$ models their measurement as the unitary $\mathcal{M}_{\text{unitary}}^{A_i}$ and when set to $x_i = 1$ models the measurement as the unitary $\mathcal{M}_{\text{unitary}}^{A_i}$ followed by a set of projectors associated with the measurement outcome a_i .

In the present case, we have four agents. In the full augmented circuit, we would, therefore, have four setting variables, one for each agent. Moreover, in the FR protocol, Ursula and Wigner announce their classical outcomes u and w in each run of the protocol and halt when they obtain $u = w = \text{ok}$. Therefore, in the full augmented circuit this communication channel between Ursula and Wigner which captures this announcement will also be present. However, for the purpose of our discussion here and to see the resolution of the paradox, it suffices to work with a much simplified augmented circuit, where we fix Ursula’s and Wigner’s settings to 1 (as they are effectively classical from the perspective of all agents involved) [32] and ignore the explicit communication channel for the announcement as this is a postprocessing. Then, we need only to assign settings x_1 and x_2 to the agents Alice and Bob.

We can now write the enlarged set of projectors $\Pi_{x_1}^A$ and $\Pi_{x_2}^B$ [cf. Eq. (9)] as

$$\begin{aligned}\Pi_0^A &:= \{\pi_{0,\perp}^A = 1_{\text{RA}}\}, \\ \Pi_1^A &:= \{\pi_{1,0}^A = |00\rangle\langle 00|_{\text{RA}}, \pi_{1,1}^A = |11\rangle\langle 11|_{\text{RA}}\}, \\ \Pi_0^B &:= \{\pi_{0,\perp}^B = 1_{\text{SB}}\}, \\ \Pi_1^B &:= \{\pi_{1,0}^B = |00\rangle\langle 00|_{\text{SB}}, \pi_{1,1}^B = |11\rangle\langle 11|_{\text{SB}}\}.\end{aligned}\quad (10)$$

These capture the fact that, when Alice’s setting $x_1 = 0$, her measurement is modeled as a unitary evolution and the a is set deterministically to the trivial value $\perp$. When $x_1 = 1$, Alice’s measurement is first modeled unitarily, and then her measurement outcomes $a = 0$ and $a = 1$ are identified by the projectors $\pi_{1,0}^A$ and $\pi_{1,1}^A$. The case for Bob is similar.

With this, we obtain the (simplified) augmented circuit for the entanglement version of the FR protocol which is illustrated in Fig. 5. Note that here Alice and Bob act at the same time $t = 2$ and Ursula and Wigner act at $t = 3$, but we could always have the agents’ operations occur at different times without affecting the circuit structure, such that this circuit fits within the general form in Fig. 3. We do not do this here for simplicity, but it is easy to see that this transformation would not affect any of the arguments.

B. Explicit version of the statements that resolve the paradox

As reviewed in Appendix D 2, the four logical statements involved in the entanglement version of the FR paradox (see Fig. 4 for a quick recap of the protocol) are as follows:

$$\begin{aligned}u &= \text{ok} \wedge w = \text{ok}, \\ u &= \text{ok} \Rightarrow b = 1, \\ b &= 1 \Rightarrow a = 1, \\ a &= 1 \Rightarrow w = \text{fail}.\end{aligned}\quad (11)$$

The paradox ensues since the last three statements can be combined using classical logic to yield $u = \text{ok} \Rightarrow w = \text{fail}$, which contradicts the first statement. These statements follow due the following four conventional predictions (Definition 6) of the four-agent EWFS specified by this protocol while postselecting on an experimental run where $u = w = \text{ok}$ is obtained by the superagents Ursula and Wigner. The first prediction in the list below guarantees this postselection will eventually succeed:

$$\begin{aligned}P_{\text{conv}}(u = \text{ok}, w = \text{ok}) &= \frac{1}{12}, \\ P_{\text{conv}}(b = 1 | u = \text{ok}) &= 1, \\ P_{\text{conv}}(a = 1 | b = 1) &= 1, \\ P_{\text{conv}}(w = \text{fail} | a = 1) &= 1.\end{aligned}\quad (12)$$

From our main theorem, Theorem C1, it follows that these conventional predictions are equivalent to the following four setting conditioned predictions (that can be computed from the augmented circuit in Fig. 5):

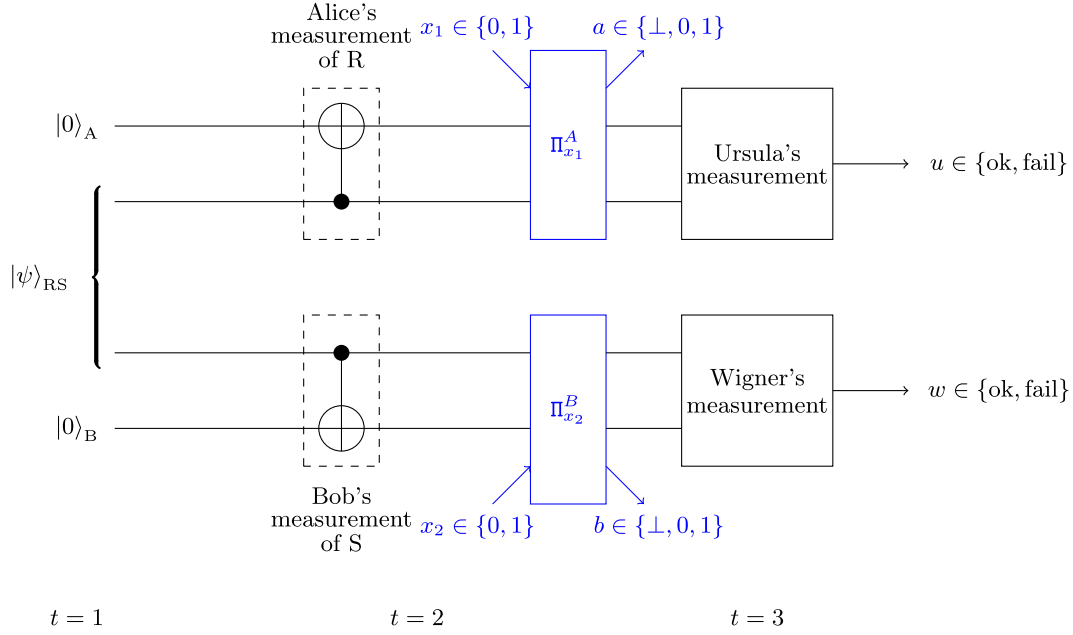


FIG. 5. Augmented circuit for the entanglement version of the FR protocol. Unlike the circuit in Fig. 4, this circuit allows agents to model the measurements of Alice and Bob as unitary evolutions and at the same time also reason about their classical measurement outcomes. To model Alice unitarily, one must set $x_1 = 0$, in which case the trivial outcome $a = \perp$ is obtained deterministically and the blue box acts as an identity operation. To reason about Alice's classical outcome, one must set $x_2 = 1$, in which case the blue box applies the projectors corresponding to the classical outcome $a \in \{0, 1\}$. All possible reasoning in this FR setup can be derived from this circuit under different choices of settings, but these statements can no longer be combined to yield a paradox as explained in the main text.

$$\begin{aligned}
 P(u = \text{ok}, w = \text{ok} | (x_1, x_2) = (0, 0)) &= \frac{1}{12}, \\
 P(b = 1 | u = \text{ok}, (x_1, x_2) = (0, 1)) &= 1, \\
 P(a = 1 | b = 1, (x_1, x_2) = (1, 1)) &= 1, \\
 P(w = \text{fail} | a = 1, (x_1, x_2) = (0, 1)) &= 1.
 \end{aligned} \tag{13}$$

The three logical statements associated with the last three setting-conditioned predictions, together with the corresponding original three logical statements of FR, are given in Table I for comparison.

Then, we can immediately see that, while the original statements can be chained together using classical logical

TABLE I. Explicit versions of the FR statements as given in our framework, which provides a logically consistent resolution to the FR apparent paradox without giving up any of the original assumptions of FR. While the original statements can be chained together to yield $u = \text{ok} \Rightarrow w = \text{fail}$, which yields a contradiction along with the fact that $P(u = w = \text{ok}) > 0$, our explicit version of the statements cannot be chained together to yield this conclusion even using the axioms of classical logic.

Original formulation	Our explicit formulation
$u = \text{ok} \Rightarrow b = 1$	$(x_1, x_2) = (0, 1) \wedge u = \text{ok} \Rightarrow b = 1$
$b = 1 \Rightarrow a = 1$	$(x_1, x_2) = (1, 1) \wedge b = 1 \Rightarrow a = 1$
$a = 1 \Rightarrow w = \text{fail}$	$(x_1, x_2) = (1, 0) \wedge a = 1 \Rightarrow w = \text{fail}$

rules to yield $u = w = \text{ok} \Rightarrow w = \text{fail}$ [Eq. (D7)], the explicit version of the statements obtained in our framework cannot be chained together in the same manner, even under the standard rules of classical logic.

The fact that the probabilities of Eqs. (12) and (13) indeed match can be seen by explicitly writing out the expression of the conventional prediction using the Born rule, which we show in full detail in Appendix G 1. We illustrate this here for some of these cases. Consider the conventional prediction $P_{\text{conv}}(u = \text{ok}, w = \text{ok})$. This is computed in the circuit in Fig. 4; i.e., applying the two CNOT gates to the initial state of $|\psi\rangle_{\text{RS}}$ and memories initialized to $|0\rangle_{\text{A}}$ and $|0\rangle_{\text{B}}$, one obtains $|\psi\rangle_{\text{RASB}} = (1/\sqrt{3})(|0000\rangle + |1100\rangle + |1111\rangle)_{\text{RASB}}$. The measurements of Ursula and Wigner act on this state. Indeed, our augmented circuit in Fig. 5 for the case $(x_1, x_2) = (0, 0)$ is equivalent to the circuit in Fig. 4. Thus writing out the two predictions, we have

$$\begin{aligned}
 P_{\text{conv}}(u = \text{ok}, w = \text{ok}) &= P(u = \text{ok}, w = \text{ok} | (x_1, x_2) = (0, 0)) \\
 &= |\langle \text{ok} |_{\text{RA}} \otimes \langle \text{ok} |_{\text{SB}} | \psi \rangle_{\text{RASB}}|^2.
 \end{aligned} \tag{14}$$

For comparison, consider the conventional prediction $P_{\text{conv}}(a = 1 | b = 1)$. This is fully specified by $P_{\text{conv}}(a = 1, b = 1)$, and it will be more illustrative to consider that. This is obtained by applying the

computational basis measurement on R and on S to the initial state $|\psi\rangle_{\text{RS}}$, which is equivalent to applying the $\{|00\rangle, |11\rangle\}$ basis measurement to $|\psi\rangle_{\text{RASB}}$. This is exactly what one would get when using the settings $(x_1, x_2) = (1, 1)$ in the augmented circuit in Fig. 5. We have

$$\begin{aligned} P_{\text{conv}}(a = 1, b = 1) \\ &= P(a = 1, b = 1 | (x_1, x_2) = (1, 1)) \\ &= |\langle 11 |_{\text{RA}} \otimes \langle 11 |_{\text{SB}} \cdot |\psi\rangle_{\text{RASB}}|^2. \end{aligned} \quad (15)$$

The equivalence of the remain predictions can be similarly shown, as detailed in Appendix G 1.

Remark 3 (on the role of the projection postulate). As seen in Sec. II, Wigner’s friend scenarios generally involve three aspects of quantum theory: unitary evolution, projection postulate, and the Born rule. While all three aspects are present in textbook quantum theory, one may, however, question the role of the projection postulate in arguments based on measurement probabilities.

In the entanglement version of the FR paradox considered here, the apparent contradiction arises from combining the logical statements in Eq. (11), which directly follow from the standard predictions in Eq. (12). These predictions (measurement probabilities) are derived using the Born rule and unitary modeling of measurements, without requiring the projection postulate, which specifies the postmeasurement state. Importantly, our resolution, based on the setting-conditioned predictions in Eq. (13), also operates entirely at the level of measurement probabilities and does not invoke the postmeasurement state. Therefore, the resolution given in Table I also need not invoke the projection postulate.

An important but subtle point here is that, even though the setting-conditioned predictions involve the setting 1 (which was described in the augmented circuit as associated with the state update of the projection postulate), one does not require this state update rule for computing the predictions. Nevertheless, the projectors associated with the setting 1 description are *necessary* to identify the measurement outcome and compute its probability via the Born rule. Moreover, they specify the measurement basis.

For example, to compute the joint probability of Ursula and Bob’s outcomes u and b , Alice’s measurement is modeled unitarily (setting $x_1 = 0$), and its basis is irrelevant for the prediction. In contrast, Bob’s measurement, described under setting $x_2 = 1$, requires the basis projectors $\pi_{1,0}^{\text{B}} = |00\rangle\langle 00|_{\text{SB}}$, $\pi_{1,1}^{\text{B}} = |11\rangle\langle 11|_{\text{SB}}$ from Eq. (10) to identify b and compute its probability. Thus, setting 1 need not imply an objective “collapse” but rather encodes the information required to specify and interpret a classical record, independently of any postmeasurement state. This can be perfectly consistent with a unitary description of the measurement by another agent, for whom the original measurement is regarded as an evolution of a closed system

and the classical record is unknown (associated with trivial outcome $\perp$).

This illustrates that even versions of the FR argument that avoid invoking the projection postulate or collapse dynamics can be resolved within our framework, simply by carefully accounting for the knowledge and conditioning used in the reasoning.

C. Setting dependence: A refined interpretation of the FR paradox

In Sec. IV C, we identified a new assumption **I** (setting independence) and showed that inconsistent quantum predictions in EWFSs arise only when assuming **I** in a situation where predictions do depend on the setting. Analyzing this assumption for the FR scenario sheds light on the root cause of such FR paradoxes yielding more refined physical interpretation.

In Appendix G 3, we show the setting dependence of predictions in the FR scenario (i.e., a violation of **I**), by explicitly computing them. For instance, it follows from the detailed analysis of Appendix G 1 that the conventional prediction $P_{\text{conv}}(w = \text{fail} | a = 1)$ (used to derive the statement $a = 1 \Rightarrow w = \text{fail}$) is given as

$$\begin{aligned} P_{\text{conv}}(w = \text{fail} | a = 1) \\ &= P(w = \text{fail} | a = 1, (x_1, x_2) = (1, 0)) = 1. \end{aligned} \quad (16)$$

$P(w = \text{fail} | a = 1, (x_1, x_2) = (1, 0))$ is indeed setting dependent (violates **I**), as one can verify (see Appendix G 3) that

$$P(w = \text{fail} | a = 1, (x_1, x_2) = (1, 1)) = \frac{1}{2} \neq 1. \quad (17)$$

Similarly, the setting dependence of other predictions of the FR scenario can also be verified in our framework.

Using the general results of Sec. IV and the above analysis of the FR scenario, we can readily prove the following theorem regarding our assumptions applied to the FR scenario. This applies to both the entanglement version (that was the focus here) and the prepare-and-measure versions (described in Appendix G 2).

Theorem 1. There exists a consistent description of the FR protocol (both versions) that satisfies all five assumptions **Q**, **U**, **C**, **D**, and **S** but violates **I** for certain logical setting-conditioned predictions. Furthermore, when simultaneously assuming **Q**, **U**, **C**, **D**, and **S** in the FR protocol, additionally imposing **I** on at least one logical setting-conditioned prediction is a necessary condition for reproducing the apparent FR paradox, while imposing **I** on all logical setting-conditioned predictions is a sufficient condition for the same.

FR have claimed that their assumptions Q , U , C , D , and S lead to a contradiction in a physical theory that reproduces the quantum predictions of the FR scenario, while we

have formalized a version **Q**, **U**, **C**, **D**, and **S** of these, showing that they can always be applied consistently even while reproducing the FR predictions. This calls for closer examination of FR's claim to understand this apparent mismatch.

FR's work suggests that their assumptions **Q**, **U**, **C**, **D**, and **S** should be interpreted as capturing the validity of unitary "quantum theory" and "classical logic" applied to the knowledge of agents. However, the formalization of the FR assumptions leave ambiguities in the definition of "agents' knowledge" and in how agents model measurements in each statement they make. Indeed, we have provided a formalization of the same principles via **Q**, **U**, **C**, **D**, and **S**, which provably avoids FR-type paradoxes, and have shown that the additional assumption **I** (setting independence) must be imposed to recover an apparent contradiction.

This contrast necessitates recognizing two versions of quantum theory in the context of EWFSs: (i) *Heisenberg-cut-independent* and (ii) *Heisenberg-cut-dependent* versions.

In essence, FR's result can be interpreted as revealing a contradiction between version (i) of quantum theory and classical logic within a specific EWFS, while our findings establish the general consistency between version (ii) of quantum theory and classical logic across all EWFSs. Then, the two sets of results are mutually consistent, even though our results imply the broader claim of FR, regarding the inconsistency of quantum theory does not hold.

Moreover, the violation of **I**—the dependence of predictions on Heisenberg cuts in EWFSs—is not unexpected once the concept of such cuts is formalized in terms of different channels describing a measurement. Not only is the **I** assumption violated in the FR scenario as shown here, this is also the case in Wigner's original thought experiment. Indeed, Wigner clearly points to this effect in the original paper where the thought experiment was introduced, although evidently not in the language of the settings we use here. As we can see from the review of Wigner's experiment in Sec. II, the core message is that the ambiguity in how a measurement is modeled, in light of the unitarity vs projection postulates (which in our framework is labeled by the settings), does have empirical consequences in Wigner's scenario.

This highlights that in EWFSs, when assuming the universal validity of unitary quantum theory, it is natural to consider version (ii) of quantum theory, which incorporates the Heisenberg-cut dependence of predictions. This version can be formalized while remaining perfectly consistent in all EWFSs as we have shown. Both Wigner's original result as well as FR's result, therefore, serve to raise caution on the dangers of using version (i) of quantum theory in such EWFSs. However, FR's arguments differ from Wigner's as FR arrive at the apparent contradiction while requiring only agents performing compatible

measurements to reason about each other in each statement [33], while the apparent paradox obtained in Wigner's experiment requires a superagent (Bob) to issue statements about an agent (Alice) who performs an incompatible measurement.

D. Link to (non)absoluteness of observed events

Although FR's arguments center on agents' reasoning, our analysis here takes a step back and focuses on the more fundamental aspect of the predictions of quantum theory for the scenario. This allows us to draw insights on the failure of an absolute notion of observed events in EWFS (as highlighted in Refs. [6,19]) and its relation to the settings we have introduced in this work.

AoE entails that the predictions of the scenario can be derived from a single joint probability distribution on the observed (nontrivial) outcomes of all agents. However, the quantum predictions of the FR scenario given in Eq. (12) are not compatible with a single joint distribution $P(u, w, a, b)$ on the observed outcomes of all four agents.

We have shown that these conventional predictions of Eq. (12) are equivalent to the setting-conditioned predictions of Eq. (13). The setting choices are present but not made explicit in the conventional representation.

Once we account for the setting dependence of these predictions using their explicit form given in Eq. (13), their incompatibility with a single, well-defined joint distribution $P(u, w, a, b)$ independent of any settings becomes immediate. This is because those predictions arise from different conditional probability distributions but by ignoring the conditioning information (on which the prediction depends); therefore, whenever there is such setting dependence in a scenario (violation of **I**), we cannot expect AoE to hold for the corresponding conventional predictions of that scenario.

Because of space considerations, we leave a detailed discussion of other no-go results for EWFSs relating to the absoluteness of events [6,7], particularly the LF theorem [7], and the relation between AoE and **I** to a follow-up paper [22]. However, we note that our **I** assumption plays a very different role in the FR vs LF scenarios. While we have shown that the violation of **I** is sufficient to fully evade the conclusion of FR's paper that "quantum theory cannot consistently justify the use of itself," the violation of **I** in the LF scenario cannot be used to evade their conclusions; rather, it sheds deeper light on the structure and meaning of their absoluteness of events assumption.

VI. EMERGENCE OF ABSOLUTE MEASUREMENT EVENTS

In Wigner's friend scenarios, relational approaches typically propose to evade no-go theorems by relaxing the objectivity of measurement outcomes and demanding that measurement outcomes are to be defined relative to an

agent, a context, a world, or some other new concept introduced within the framework (see also the discussions in Refs. [3,4,11]). However, this leaves open the crucial question of how one can recover predictions and observations of existing real-world quantum experiments where measurement outcomes and probabilities are objective for all agents and do not appear to be relational for any given protocol.

In this section, we address this question and show how our formalism recovers the apparent objectivity of measurement outcomes and the standard predictions of quantum theory in existing real-world experiments, even though it provides a fundamentally relational and consistent measurement model for general EWFSs where an absoluteness notion of measurement events may not exist (as discussed in the preceding section).

A. Criteria for superagency: Distinguishing standard and genuinely Wigner's friend scenarios

To address the open question motivated above, we first develop formal criteria to distinguish genuinely Wigner's-friend-type scenarios from standard quantum scenarios where agents do not measure each other's memories or labs in a nontrivial manner (or, more colloquially, when they do not “Hadamard each others' brains”).

Recall that the memory M_i of each agent A_i plays the role of “their entire lab except the system that they measure”; i.e., the quantum system S_i which A_i measures together with their memory M_i constitute a minimal model of A_i 's lab needed for EWFS arguments. We now proceed to formally define what is meant by “acting nontrivially on an agent's memory” and thereby establish a formal criterion distinguishing when one agent does not act as a superagent to another in an EWFS. Considering this criterion for all pairs of agents in the EWFS, we arrive at a concept that we term as the *nonsuperagent structure* ($n\mathcal{SA}$) of the EWFS, which will allow one to differentiate parts of a general EWFS as standard quantum sectors as opposed to genuine Wigner's friend experiments.

Here, it is important to note that, even in everyday scenarios, agents do act on the memories of each other in the sense that an agent can consult their memory, which stores a classical outcome, and communicate it to another agent. This may influence the operations and reasoning process of the second agent. Therefore, we need a definition that is not too restrictive to forbid this kind of “trivial” or “standard” way of acting on each others' memories while still strong enough to identify “nontrivial” or “nonstandard” ways in which the operation performed by one agent in a Wigner's friend scenario can act on the memory or lab of another agent.

A key preliminary concept required for developing this is the operational equivalence between two EWFSs.

Definition 12 (operationally equivalent EWFSs). We say that two N -agent EWFSs involving sets $\{A'_1, \dots, A'_N\}$ and

$\{A_1, \dots, A_N\}$ of agents are operationally equivalent if and only if the following hold.

- (i) There is one-to-one identification between the systems $S' = \{S'_1, \dots, S'_m\}$ and $S = \{S_1, \dots, S_m\}$, agents $\{A'_1, \dots, A'_N\}$ and $\{A_1, \dots, A_N\}$, memories $\{M'_1, \dots, M'_N\}$ and $\{M_1, \dots, M_N\}$, measurements $\{\mathcal{M}^{A'_1}, \dots, \mathcal{M}^{A'_N}\}$ and $\{\mathcal{M}^{A_1}, \dots, \mathcal{M}^{A_N}\}$, subsets $S'_i \subseteq S' \cup M' \setminus \{M'_i\}$ and $S_i \subseteq S \cup M \setminus \{M_i\}$ of systems on which each measurement acts nontrivially, and sets of operations $\{\mathcal{E}'_1, \dots, \mathcal{E}'_N\}$ and $\{\mathcal{E}_1, \dots, \mathcal{E}_N\}$, with outcome sets O'_i and O_i being equivalent.
- (ii) The augmented circuits of the two EWFSs yield the same setting-conditioned predictions; i.e., for all disjoint subsets $\vec{a}_j$ and $\vec{a}_l$ of outcomes in one EWFS and corresponding subsets $\vec{a}'_j$ and $\vec{a}'_l$ of outcomes in the other, as well as all settings $\vec{x}$ in one and corresponding settings $\vec{x}'$ in the other, we have

$$\begin{aligned} P(\vec{a}_j = \vec{a}'_j | \vec{a}_l = \vec{a}'_l, \vec{x} = \vec{x}') \\ = P(\vec{a}'_j = \vec{a}_j | \vec{a}'_l = \vec{a}_l, \vec{x}' = \vec{x}). \end{aligned} \quad (18)$$

The formal definition of the nonsuperagent structure $n\mathcal{SA}$ is given in Appendix C (Definition C8) as this combines the above definition with those of causal structure and directed paths from Appendix C 1. Below, we provide a detailed and illustrative intuition for the definition which suffices to follow our main results.

The main idea behind the $n\mathcal{SA}$ is that, for an agent A_j not to act as a superagent to another agent A_i , it suffices to know either that A_j does not act after A_i in the causal structure (in which case they do not act on A_i 's memory at all [35]) or that all operations in the augmented circuit occurring after A_i 's measurement, including A_j 's operations, can be simulated by equivalent operations that act trivially on A_i 's memory.

A simple example illustrates this concept, showing that it does not exclude usual scenarios where agents may communicate different information based on the outcomes stored in their memories. Notice that the postmeasurement state of an agent's memory M_i and measured system S_i after a measurement $\mathcal{M}^{A_i}$ is symmetric in the exchange of M_i and S_i for both setting choices $x_i = 0$ and $x_i = 1$, as shown in Eqs. (3) and (7). This symmetry exists because the memory is perfectly correlated with the system in the measurement basis, obtained by coherently or incoherently copying (depending on the setting) the system state in that basis. Therefore, any scenario where an agent prepares a new state for another agent based on the state stored in their memory can be perfectly mimicked by an equivalent operation that prepares the new state based on the state of the system, acting trivially (as an identity) on the memory.

However, a scenario where Wigner performs a Hadamard operation on a friend's brain, or undoes a friend's measurement, involves a nontrivial joint operation on the memory and system that cannot be emulated by an operation acting on the system alone. Therefore, when A_i acts before A_j in the causal order, the key property distinguishing whether or not A_j acts as a superagent to A_i is whether A_j 's operations necessarily act jointly on A_i 's lab (system and memory). The nonsuperagent structure nSA of an EWFS (defined in Definition C8) includes an ordered pair (A_i, A_j) of agents whenever A_j does not act as a superagent to A_i in the EWFS.

Importantly, the nSA of an EWFS is also a physical, objective and perspective-independent property of the protocol, as is the causal structure of an EWFS as defined in Definition C1. Although the notion of operational equivalence in Definition 13 refers to settings, the property is also independent of the settings (or Heisenberg cuts) as it is required to hold for all setting choices.

For instance, in Wigner's original experiment, where the agent Wigner W can measure the lab of the friend F in the Bell basis while the friend acts only on some quantum system (that does not include Wigner's memory or lab), applying this definition, we find that $(F, W) \notin nSA$ and $(W, F) \in nSA$. This indicates that W can act as a superagent to F but not vice versa. More generally, if another agent Ursula U can "ask" Wigner about his observed measurement outcomes but does not have nontrivial quantum control over Wigner's full lab, we can simulate Ursula's operation of "asking Wigner about his outcome" as an operation that acts directly on Wigner's system (friend's lab) and not on his memory (as discussed earlier in this section). This simulation would produce the same setting-conditioned predictions in the augmented circuit as the original scenario.

Therefore, $(W, U), (U, W) \in nSA$ while $(F, U) \notin nSA$ and $(W, U) \in nSA$, and, of course, $(F, F), (W, W), (U, U) \in nSA$ since no agent acts on their own memory through the channel $\mathcal{E}_i$ that they implement after their measurement.

In the (entanglement version of the) FR protocol reviewed in Appendix D 2, we have $nSA = \{(A, B), (B, A), (U, W), (W, U), (U, B), (B, U), (W, A), (A, W), (U, A), (W, B), (A, A), (B, B), (U, U), (W, W)\}$, but this is not a standard quantum experiment, as the pairs (A, U) and (B, W) do not appear in nSA ; the latter agent acts as a superagent to the former. The nSA of the prepare-and-measure version of the FR protocol is identical.

We are now ready to formally define what constitutes a standard quantum experiment.

Definition 13 (standard quantum scenario). An EWFS with N agents $\{A_1, \dots, A_N\}$ is said to correspond to a standard quantum scenario if $(A_i, A_j) \in nSA$ for all $A_i, A_j \in \{A_1, \dots, A_N\}$. In case this holds for a subset of the first k agents $\{A_1, \dots, A_k\}$ (who act at time steps $t_1 < \dots < t_k$)

in the context of a larger N -agent EWFS, we call $\{A_1, \dots, A_k\}$ a standard quantum sector of the EWFS. Otherwise, we call it a nonstandard scenario or sector accordingly.

For instance, in FR's protocol, the agents Alice and Bob who act first form a standard quantum sector, but once we include the superagents Ursula and Wigner, the sector is no longer standard.

Nonstandardness as a signature of genuine Wigner's friendness?—An interesting question is whether scenarios that are nonstandard according to our Definition 13 can be considered as exhibiting a genuinely Wigner's-friend-type aspect. One intuition in favor of this is that, by construction, nonstandardness captures that there is at least one pair of agents such that one (say, A_j) acts on the memory of another (say, A_i) in a nontrivial manner that cannot be regarded as A_j simply "asks" A_i their outcome. However, this point needs to be further investigated at both a conceptual and technical level to make conclusive statements; after all, there can be different equally well-motivated criteria for "genuineness" of a nonclassical resource, as the vast literature on quantum nonlocality and entanglement highlights.

Our formalism provides a first consistent, general, and fully formal platform for formulating and investigating such questions for EWFS. This gives the potential to understand, in terms of fundamental principles and axioms, the quantum resource associated with Wigner's-friend-type no-go results that fundamentally distinguish them from no-go results (such as Bell's theorem) where agents are not treated quantum mechanically. We leave this for future work.

B. Recovering Heisenberg-cut independence in standard quantum experiments

In the previous sections of the paper, we have shown that the choice of settings (which provide a circuit-based formalization of the choices of Heisenberg cuts) do affect the empirical predictions in Wigner's original scenario as well as its extensions such as FR and LF. That is, the **I** assumption (setting independence) is violated here. This differs from our standard intuitions and usage of quantum theory to describe realistic experiments, where we do not have to consider any such settings or Heisenberg cuts and where observed outcomes appear to be objective records independent of any such concepts.

In order to show that our framework correctly reproduces the known predictions and observations of standard quantum experiments conducted so far, we must show that the setting variables can be safely dropped from the predictions and the augmented circuit when analyzing such experiments. The following results formalize this intuition concretely using our definition of standard quantum scenarios. At a foundational level, this will shed light on how the perceived objectivity or nonrelationalism of observed

measurement events emerges within standard quantum sectors, where the agents involved do not have full quantum control over each other's labs or memories.

Theorem 2 (nonaction on memory and setting independence). Consider an EWFS and a subset A_K of agents therein. Suppose that $A_i \notin A_K$ is another agent in the EWFS such that no agent in A_K acts as a superagent to A_i i.e., $(A_i, A_k) \in n\mathcal{SA} \ \forall \ A_k \in A_K$. Then, for every partition $A_K = \{A_{j_1}, \dots, A_{j_p}\} \cup \{A_{l_1}, \dots, A_{l_q}\}$ of A_K , the setting-conditioned prediction $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi})$ is independent of the setting x_i ; that is,

$$\begin{aligned} P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi}) \\ = P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} \setminus x_i = \vec{\xi} \setminus \xi_i), \end{aligned} \quad (19)$$

where $\vec{x} \setminus x_i = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_N)$ is the setting vector with the i th component omitted, and similarly for the vector of setting values $\vec{\xi} \setminus \xi_i$. This expression is equivalent to the conditional independence given in Eq. (J10).

A proof of the above theorem can be found in Appendix J. As a corollary of Theorem 2, we can immediately recover the full setting independence of predictions in standard quantum experiments.

Corollary 4 (full setting independence in standard quantum theory). In any EWFS corresponding to a standard quantum scenario, every nontrivial setting-conditioned prediction is setting independent. Formally, for any disjoint sets $\vec{a}_j$ and $\vec{a}_l$ of outcomes in the EWFS, $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi})$ is independent of settings x_i for all $i \in \{j_1, \dots, j_p, l_1, \dots, l_q\} := \mathcal{JL}$; i.e., Eq. (J6) holds for all such i . Specifically, in such standard scenarios, all nontrivial predictions, i.e., those where $a_i \neq \perp$ for all $i \in \mathcal{JL}$, can equivalently be expressed in a fully setting-independent manner as given below:

$$\begin{aligned} P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi}) \\ = P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, x_i = 1, \ \forall \ i \in \mathcal{JL}) \\ := P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l). \end{aligned} \quad (20)$$

Going beyond predictions and to the underlying circuit representation, we expect that standard quantum experiments involving some measurements and channels can be represented in terms of quantum circuits with no ambiguity in how a measurement is modeled (i.e., circuits with no setting variables for measurements). Here, two equivalent types of circuit representations are possible, which are both commonly used within the standard quantum computing paradigm: Either we can model each measurement as acting only on the measured system and yielding a (nontrivial) classical outcome at the time of the measurement, or we can model each measurement as a unitary interaction between a system and ancilla at the time of measurement and probabilities can be extracted by

measuring all the ancillas at a later time at the end of the experiment.

We leave a formal definition of these to Appendix F, where we prove that, for EWFSs corresponding to standard quantum scenarios, the augmented circuit can equivalently be reduced to either of these expected forms. These two types of standard quantum circuits are illustrated in Figs. 7 and 8 in the same appendix.

VII. DISCUSSION

A. Sound scientific reasoning in EWFSs and analogies to classical multiagent reasoning

The results of Sec. III guarantee that our formalism allows quantum agents to make predictions and to reason consistently about experiments and each other's knowledge, even if unitary quantum theory is universally valid. Here, we illustrate a general paradigm for scientific reasoning indicated by our results, which allows one to straightforwardly extend our resolution to multiagent scenarios with partial knowledge. We do so while contrasting quantum and classical aspects and showing how consistency questions in EWFSs reduce to analogous issues in classical multiagent reasoning (see also Ref. [36] for comparisons between Wigner's friend and classical paradoxes).

1. Genuinely quantum and relational aspects

Our results (Definition 6 and Theorem C1) provide a concrete default rule for choosing settings to compute predictions, which makes them equivalent to conventional quantum predictions for EWFSs used in the literature while avoiding paradoxes by making the conditioning on settings explicit. Operationally, the rule says that, when computing the outcome probabilities of a subset of measurements, exactly those measurements are treated as producing "classical" records (setting 1) and the rest are modeled purely unitarily (setting 0), which can be interpreted as a choice of Heisenberg cuts (outside vs inside). When agents reason about others, the chosen cuts depend on the agent who is reasoning and the agents whose classical outcomes they are reasoning about: When A reasons about B 's outcome (and not C 's), then A will place A 's and B 's measurements on the classical side of the cut and C 's on the quantum side; but when A reasons about C 's outcome, C 's measurement will be placed on the classical side. This captures relational, nonabsolute events and the setting (Heisenberg-cut) dependence of predictions. Moreover, genuinely WF-like scenarios require nontrivial superagent actions, often linked to measurement complementarity. This suggests a nonclassical resource tied to "genuinely WF-like aspects" yet to be fully characterized in future work (see also Secs. VI A and VIII).

2. Aspects reducible to classical issues

By formalizing Heisenberg cuts as channel choices in a circuit, classical probability and logic already dictate conditioning on these choices unless (i) all agents employ the same fixed choices or (ii) the choices are proven irrelevant. Even classically, two physicists who assume different noise models N_A and N_B for the same known channel (and applied to the same input state) would generally reach different predictions unless they communicate and condition on their assumptions, and they could obtain FR-like apparent inconsistencies if they ignore such parameters (see Appendix H for a classical scenario that reproduces certain aspects of the FR paradox). On the other hand, the conclusions are independent of whether the analysis was done during the day or night; hence, one can safely ignore this information.

3. General reasoning paradigm

These examples yield a simple recipe for sound scientific reasoning in any multiagent setting:

- (1) Identify common knowledge vs variable parameters that are not common knowledge to all agents;
- (2) drop parameters shown to be redundant;
- (3) fix the remaining parameters (according to some well-defined rules) and condition on them;
- (4) test choices against data; update if falsified.

In our framework from Sec. III, the EWFS protocol description (Definition 1) is common knowledge of all agents; the only variables outside common knowledge are the settings $\vec{x}$ (the case of partial knowledge of the protocol is discussed shortly). Theorems 2 and C1 give criteria—via operational causal structure and the nonsuperagent structure—for when settings are redundant; the default rule (entailed by Definition 6 and Theorem C1) then fixes the rest and conditions the predictions on them to ensure consistency per Definition 5.

Finally, our discussion (Sec. VIIC and Appendix I) shows how setting-conditioned predictions can be falsified. The default rule for choosing the settings assumes universal validity of unitary quantum theory; if the existence of an objective-collapse mechanism is confirmed in some future loophole-free experiment, it would falsify predictions made under universal quantum theory and mandate updating certain settings from 0 to 1.

4. Efficiency of reasoning

Computing setting-conditioned predictions requires applying the Born rule to the same-size circuit as the underlying protocol, i.e., no worse than standard quantum reasoning. Processing settings adds no asymptotic overhead: One reads independences from the causal and non-superagent structures (these are precomputable) [37]. The only extra cost is communicating or storing a short binary vector of nonredundant settings—already necessary in

analogous classical ambiguities. Our framework also satisfies **C**: Communicated setting-conditioned statements need not be recomputed by recipients, while relationality appears through agents choosing different settings under the default rule.

5. Scope of our resolution and partial knowledge

We assumed that the EWFS protocol (cf. Definition 1) is common knowledge of the agents involved, in order to isolate what is genuinely WF-like. Extending to cases where agents have partial knowledge of the protocol is directly analogous to how this would be handled in classical multiagent reasoning and is based on the same general reasoning paradigm outlined above: Fewer elements of common knowledge mean more variables to condition on to avoid inconsistencies, not new quantum-specific pitfalls. The same rules apply; we omit the (straightforward) explicit algorithm.

6. Status of proposed reasoning challenges

Our work fully resolves FR-type paradoxes arising from the genuinely nonclassical aspect of AoE violation (Sec. V and Appendix G 2), and suggests that any remaining logical challenges for quantum agents in EWFSs (including partial-knowledge scenarios) are analogous to classical challenges or inconsistencies arising from failing to properly condition on assumptions outside common knowledge. For instance, Ref. [21] proposed such a challenge, inspired by the FR arguments. However, their arguments, like FR, involve ambiguities about Heisenberg cuts, measurement-channel modeling, and knowledge conditioning in computing probabilities. Similar ambiguities cause inconsistencies even classically (see also Appendix H for analogies to classical paradoxes arising from ignoring conditioning on relevant knowledge). We believe that our work, therefore, also fully resolves the challenge of Ref. [21]. Indeed, to our knowledge, there exists no EWFS counterexample (including partial-knowledge scenarios) that entails a paradox which cannot be resolved by applying our framework and general reasoning paradigm.

B. Interpretations of quantum theory

We developed our framework in three levels of generality which enable flexibility across interpretations and conceptual viewpoints.

- (i) A general EWFS formalism that allows for different rules for computing probabilities $P(\vec{a}_j|\vec{a}_l, k)$ of some outcomes $\vec{a}_j$ and $\vec{a}_l$, given certain parameters k describing the assumptions made about the scenario (such as its states and channels, etc.).
- (ii) Specializing these rules such that k refers to the settings $\vec{x}$ of the augmented circuit still leaves freedom to choose the setting values. Different interpretations suggest different k and settings in

the above two levels, e.g., Bohmian mechanics can include the hidden variables λ in the k parameters; collapse theories would fix $x_i = 1$ for all measurements, while many worlds would allow $x_i = 0$ (a purely unitary description of measurements). Our consistency theorem, Theorem C1, applies to *all* setting choices, so logical and probabilistic consistency is achievable across different interpretations. We discuss how many-worlds, collapse, hidden-variable theories, relational quantum mechanics, and QBism could consistently apply our framework in Appendix I.

- (iii) At the last level, we prescribe reasoning rules assuming universal quantum theory (this excludes some interpretations such as collapse theories), but they remain compatible with others—such as many worlds, QBism, or relational quantum mechanics—since applicability does not hinge on whether states or channels are ontic or epistemic. Related approaches such as consistent histories [38,39], decoherence [40], and the parsing rule of Ref. [12] are compared in Appendix I.

C. Physical interpretation of settings

Settings already play a necessary role in the quantum predictions of an EWFS, which are prerequisite for both FR- and LF-type arguments: One must specify whether a measurement is modeled as unitary or as producing classical records to apply the Born rule [41], and the “predictions of quantum theory” used in the EWFS literature exactly match particular setting-conditioned predictions (Sec. IV A). This applies to FR-type reasoning and to LF-type arguments about AoE (Sec. V); our framework makes these assumptions explicit. We discuss physical interpretations of these settings, noting that this can, in turn, depend on the interpretation of quantum theory to which one subscribes.

1. Heisenberg cuts and classical records

Assuming quantum theory is complete (no hidden variables beyond the operational description of the quantum states, measurements, and channels in a protocol or experiment), setting choices correspond to choices of Heisenberg cuts: $x_i = 0$ treats the lab M_i as purely quantum or inside the cut, and $x_i = 1$ treats it as producing a classical record (in the measurement basis) or outside the cut. This gives a relational model of quantum measurements, as the choice of Heisenberg cut or settings can generally depend on the prediction being considered or the perspective of an agent (as also discussed in Secs. V and VII A). This remains consistent with the universal validity of unitary quantum theory; $x_i = 1$ does not mean an objective collapse: Projectors enter only to compute probabilities (via the Born rule) about the referenced classical record (see also Remark 3).

2. Falsifying setting choices through experiments

Every EWFS equivalently specifies a unique augmented circuit (Sec. III). The settings in the augmented circuit do not correspond to choices of physical operations actually performed in a protocol, but (i) if unitary quantum theory is universally valid, they encode choices and assumptions—such as how measurements are modeled—that enter predictions and agents’ reasoning (these choices can be relational and agent dependent); or (ii) in theories that reject universal unitary evolution, they encode fundamental (often absolute for all agents) macroscopic dynamics of the theory. For example, in Wigner’s original setup, collapse models imply $x_A = 1$ for the friend (corresponding to a collapse mechanism producing classical records), while many worlds suggests $x_A = 0$ (purely unitary dynamics), leading to distinct predictions for Wigner’s outcome $P(w|x_A)$. Which choice nature prefers is ultimately empirical, and the fact that these predictions differ leaves open the possibility of a future experiment that could falsify either $P(w|x_A = 0)$ or $P(w|x_A = 1)$. This highlights that setting choices do have physical and empirical consequences in EWFSs. By explicitly keeping track of the settings and their dependence, our framework provides a way to clearly test and understand which predictions in an EWFS can be falsified and under what assumptions.

3. Time dependence of setting choices and knowledge update

The falsification of predictions in an EWFS prompts an update of our knowledge about the underlying setting choices. This requires setting dependence, since setting-independent predictions would not allow any inference about the settings. While knowledge updates in classical theories concern “absolute facts of the matter” that exist but about which one may lack knowledge, in our quantum EWFS formalism the update concerns an agent’s beliefs about the existence of classical measurement records, which may no longer be absolute facts. For example, at present we can safely treat all real-world measurements as setting 1, as there is no reason to believe there currently exists a superagent capable of performing quantum operations on an entire lab, needed to erase another agent’s records. This suggests another operational interpretation: Setting 1 reflects the belief that no superagent can jointly act on all relevant quantum degrees of freedom (d.o.f.) of some agent’s (Alice’s) lab, ensuring the persistence of Alice’s measurement records in the d.o.f. not being acted upon (see also Sec. VI A). But if such “superagent” devices were developed, we would need to classify some measurements as setting 0 to explain experiments in which records can be undone (e.g., if a sufficiently loophole-free violation of an LF inequality [7] were achieved) and update previous beliefs about setting 1. This aligns with a decoherence-based interpretation, where classical records emerge from interactions with an environment. However, there is a

crucial distinction, as those interpretations do not typically allow the possibility that the environment could later be under the quantum control of a superagent or the resulting time dependence with regard to the persistence of classical records.

4. Analogy to Maxwell's demon paradox

Maxwell's demon paradox [42] in thermodynamics is resolved by clearly adopting a single perspective (demon vs macroscopic) when computing entropies, accounting for information erasure, and keeping track of which perspective was used [43,44]. Likewise, our Corollary 3 shows that, in EWFSs, making the choice of settings clear and conditioning on them (which model Heisenberg cuts and, consequently, different agents' perspectives) is key to obtaining consistent predictions and reasoning; combining statements made across different setting choices and then ignoring them by assuming **I** creates only an *apparent* paradox that is easily resolved, as per the above analogy to Maxwell's demon.

VIII. CONCLUSIONS AND OUTLOOK

A. Conclusions

Here, we provide a high-level overview of our main results, emphasizing the key conclusions (complementing the more technical summary in Sec. IB). In this work, we develop a comprehensive framework for EWFSs, overcoming a number of challenges posed by previous no-go arguments for EWFSs [3,6,7,19].

1. Full resolution of FR paradoxes and reasoning challenges in EWFSs

Our main theorems show that the framework guarantees logical and probabilistic consistency in all EWFSs and preserves causality principles while cleanly separating objective features (e.g., causal structure) from subjective ones (e.g., predictions and agents' knowledge). As a central application, this resolves all FR-type logical paradoxes in quantum theory and the challenge posed in Ref. [21], without requiring any departure from standard quantum or probabilistic theories. The resolution holds independently of one's preferred interpretation of quantum theory (cf. Appendix I) and even without invoking the projection postulate or state-update rules—while still allowing measurements to be modeled unitarily and using the standard Born rule (cf. Remark 3).

2. Identifying the root cause of paradoxes

Predictions in EWFSs change with the choice of Heisenberg cut (i.e., whether a measurement is modeled as producing a classical record or as a purely unitary evolution, captured by our Assumption I). Mixing statements made under different choices leads to apparent

contradictions and corresponds to imposing **I** in a scenario that violates it.

3. Reasoning rules for quantum agents

To avoid such paradoxes, we provide physically motivated rules for agents' reasoning that are compatible with (i) universal quantum mechanics, (ii) classical logic for recorded outcomes, and (iii) the fundamentally relational, perspective-dependent character of EWFSs. This constructively shows that quantum theory remains self-consistent in any EWFS—there is no threat to consistent scientific reasoning or to programming future quantum devices that act as agents—contrary to the broader claims made in Ref. [3].

4. Emergence of objectivity

Our framework explains how objective, Heisenberg-cut-independent predictions emerge in real-world experiments and within specific sectors of an EWFS (formalizing the idea of a “Wigner bubble” [19]), thereby accommodating the nonabsoluteness of events and relationalism in general EWFSs while remaining consistent with the objective observations seen so far.

B. Outlook

We now outline some future directions based on our work.

1. Local friendliness and nonabsolute events

The LF theorem [7] shows that no theory simultaneously satisfying the AoE, together with certain free choice and locality principles, can reproduce specific quantum predictions in an EWFS. Using our framework, we analyze the LF scenario in forthcoming work [22], highlighting the distinct role of setting or Heisenberg-cut dependence (i.e., violation of **I**) relative to the FR scenario and linking it to AoE violation. Based on this, it would be interesting to probe the possibility of characterizing novel EWFSs and no-go theorems for AoE in combination with other fundamental principles and quantum phenomena—including complementarity, contextuality, relativistic causality, and protocols where agents perform quantum operations in superpositions of causal orders [45].

2. Quantum and relativistic causality in EWFSs

Bell's theorem shows that quantum correlations challenge explanations based on classical causal models [48], if we wish to preserve relativistic causality and free choice—motivating the development of quantum causal modeling frameworks (e.g., Refs. [49–52]). The LF theorem poses even deeper challenges [19,20], even for quantum causal modeling, since these formalisms typically assume AoE. Our framework allows one to relinquish AoE yet still recovers all (possibly subjective) predictions in an EWFS

within a single objective causal structure for the protocol: The augmented circuit is acyclic and embeddable in spacetime, preserving relativistic causality. The only subjective element concerns agents' priors over settings (Heisenberg cuts). This paves the way for a relational, operational quantum causal modeling framework that does not assume AoE and remains compatible with free choice and relativistic constraints. Extending this to scenarios involving cyclic quantum circuits (and, thus, protocols with indefinite order of operations) [46,47,53–57] which have also been linked to relativistic causality [58–61] could provide a route toward a unified treatment of quantum agents, quantum causality, and spacetime structure.

3. EWFSs beyond quantum theory

EWFSs have also been formulated and studied beyond quantum theory [28,34,62]. In Refs. [28,34], agents' memories and labs were modeled within general theories, distinguishing theory-independent analogues of projective vs unitary-like (information-preserving) updates. FR-type apparent paradoxes can then arise with PR-box resources [28,63] and, more generally, in any theory admitting certain forms of contextuality [34]. Moreover, Ref. [62] shows that any theory allowing information-preserving updates, Bell nonclassicality, and a locality principle leads to AoE violations akin to the quantum LF scenario [7]. Thus, the measurement problem and FR-type paradoxes are not uniquely quantum. Generalizing our framework beyond quantum theory—building on these works—could resolve multiagent paradoxes and enable a consistent calculus for reasoning in postquantum settings while clarifying the status of AoE and the measurement problem therein.

4. Resource-theoretic characterization of EWFSs

What informational and nonclassical resources make FR- and LF-type results possible? What sets genuine Wigner's friend scenarios apart from standard experiments where agents are effectively classical? Our Theorems 2 and C1 give structural conditions relevant to such questions—under which absolute and Heisenberg-cut-independent “classical” records emerge for all agents. Our work identifies the violation of **I** by a scenario's predictions as a key signature of WF-like (as opposed to standard) quantum experiments. However, there is room to characterize necessary and sufficient conditions that depend on the properties of states and measurements (e.g., contextuality, complementarity, etc.) for violating **I** and consequently AoE (cf. Sec. VD). More broadly, the nonsuperagent structure introduced here to distinguish standard quantum scenarios from genuine Wigner's friend scenarios motivates a new research program: developing a systematic resource theory of EWFSs (as discussed in Sec. VI)—a direction that remains to be systematically explored.

ACKNOWLEDGMENTS

We thank Renato Renner, Victor Gitton, Yilè Yīng, Marina Maciel Ansanelli, Joe Renes, Eric Cavalcanti, and Nuriya Nurgalieva for insightful discussions. V. V. acknowledges support from an ETH Postdoctoral Fellowship and from the PEPR integrated project No. EPiQ ANR-22-PETQ-0007 as part of Plan France 2030. M. P. W. acknowledges funding from the Swiss National Science Foundation (AMBIZIONE Fellowship, No. PZ00P2_179914). Both authors acknowledge support from NCCR QSIT.

DATA AVAILABILITY

No data were created or analyzed in this study.

APPENDIX A: GENERALITY OF THE DEFINITION OF EWFS

Here, we discuss the justification for the generality of Definition 1 of an extended Wigner's friend scenario that we have proposed in this paper. The idea is that Definition 1 encompasses all finite multiagent quantum protocols where agents' memories (in which they store the measurement outcome) are modeled as quantum systems and where one agent can have full quantum control over the labs (measured system and memory) of other agents in the scenario. Here, the finiteness applies both to the Hilbert space dimensions and the number of information-processing steps.

Generically, we can model such scenarios by considering a set of N agents $A = \{A_1, \dots, A_N\}$ and a set of m systems $S = \{S_1, \dots, S_m\}$ under study and a set $M := \{M_1, \dots, M_N\}$ of systems, one for each agent A_i which models their memory where they store the outcomes of measurements that they perform.

For simplicity but without loss of generality, we take the lab of each agent A_i to consist of the system S_i that they measure, along with their memory M_i . Each agent $A_i \in A$ performs a measurement $\mathcal{M}^{A_i}$ on some subset $S_i \subseteq S \cup M \setminus \{M_i\}$ of systems, which can include the memories of other agents, and stores the outcome of the measurement in their memory M_i [64]. Note that S_i is a subset of systems and memories of other agents, which allows each agent to possibly act as a superagent to any subset of the other agents by measuring their memories. We have no loss of generality in assuming that each agent performs one measurement, because any scenario where one agent performs multiple measurements can be brought to this form by modeling them as multiple agents, each performing one measurement.

Next, also without loss of generality, we can assume that each agent A_i acts at a distinct time step t^i with $t^i < t^j$ for $i < j$, since any physical scenario where the same agent acts at different times can equivalently be modeled in terms of multiple agents each acting at distinct times (since our

definition allows for communication channels that carry the relevant information between the time steps). Furthermore, we can model each measurement $\mathcal{M}^{A_i}$ as acting on the whole set S of systems and the set of all memories $M_1, \dots, M_{i-1}$ of previous agents since any operation on a subset S_i of systems can be trivially enlarged into an operation on all systems by appending the identity on the complementary set of systems. Agent A_i may also perform certain fixed transformations (corresponding to quantum channel $\mathcal{E}^i$ on $S \cup M$) in between time steps t_i and t_{i+1} ; for instance, agent A_1 may measure the system $S_1 \in S$ and, depending on their measurement outcome, could perform a different transformation on some initial state of S_2 which they send to agent A_2 , who could then measure S_2 .

The circuit in Fig. 2 illustrates this general form of an EWFS that we consider, where $\{\mathcal{E}^i\}_{i=1}^N$ are fixed transformations that all agents agree on as part of the protocol while $\{\mathcal{M}^{A_i}\}_{i=1}^N$ denote the measurements, one for each agent.

Furthermore, consider that an agent A_i measures a subset S_i of the systems through a measurement $\mathcal{M}^{A_i}$. Let a_i be the random variable associated with the measurement outcome, which takes values a_i in the set $\{0, 1, \dots, d_{S_i} - 1\}$, where d_{S_i} is the Hilbert space dimension (assumed to be finite) of the system S_i measured by A_i . Without loss of generality, we can model this as a projective measurement, since any measurement on finite-dimensional systems can be purified to a projective measurement involving rank 1 projectors on a larger set of systems through Neumark dilation; see, e.g., Sec. 9-6 in Ref. [65]. Explicitly, we can use the projective measurement $\{\pi_{a_i}^{S_i} = |a_i\rangle\langle a_i|_{S_i}\}_{a_i \in \mathcal{O}_i}$, where $\{|a_i\rangle_{S_i}\}_{a_i \in \mathcal{O}_i}$ forms an orthonormal basis of S_i . While this looks like a computational basis due to the choice of outcome labels $\{0, 1, \dots, d_{S_i} - 1\}$, the choice of which basis to associate with these labels is arbitrary, and, therefore, the measurement may correspond to an arbitrary orthonormal basis.

Together with all these simplifications, we arrive at the general form of Definition 1.

APPENDIX B: DISTINGUISHING PREDICTIVE AND OBSERVATIONAL STATEMENTS: REFINING CONSISTENCY

In the main text, we used Σ to refer to the set of all statements obtained from predictions in an EWFS, in Definition 3. We motivated the need to distinguish such predictive statements from observational statements through examples of simple classical scenarios. Here, we define observational statements and discuss how our results immediately generalize to ensure consistency of predictive and observational statements together. To make the distinction clear, in this section, we explicitly write $\Sigma_{\text{pred}} := \Sigma$.

Definition B1 (observational statements). Observational statements are statements in the set $\Sigma_{\text{obs}} := \{$ “Based on

observation, I am certain that the outcomes $\vec{a}_j$ takes values $\vec{a}_j$.” $\}_{\vec{a}_j, \vec{a}_j}$. In logical notation, we denote elements of the set as $\vec{a}_j = \vec{a}_j|_{\text{obs}}$.

Definition B2 (consistency of observational statements). A set Σ_{obs} of observational statements is consistent iff $\vec{a}_j = \vec{a}_j|_{\text{obs}} \in \Sigma_{\text{obs}}$ implies that $\neg \vec{a}_j|_{\text{obs}} \notin \Sigma_{\text{obs}}$.

Definition B3 (global consistency). A set $\Sigma_{\text{pred}} \cup \Sigma_{\text{obs}}$ of predictive and observational statements obtained in an EWFS is said to be globally consistent if Σ_{pred} and Σ_{obs} are consistent according to Definitions 5 and B2, respectively, and, additionally, $S := \vec{a}_j = \vec{a}_j|_{\text{obs}} \in \Sigma_{\text{obs}}$, then $S' \in \Sigma_{\text{pred}}$, where S' is a statement associated with a prediction $P(\vec{a}_j = \vec{a}_j|k = \xi) > 0$.

Recall that within our framework the scenario parameters k are instantiated by the setting vector $\vec{x}$, which give us setting-conditioned predictions, e.g., $P(\vec{a}_j = \vec{a}_j|\vec{x} = \vec{\xi})$.

In any EWFS, a given choice of settings $\vec{x} = \vec{\xi}$ allows one to fix all the channels involved the scenario. Generically, the choice of how these setting values must be chosen is given by some set of reasoning rules $\mathcal{R}$. This rule may provide an absolute choice of setting values $\vec{x} = \vec{\xi}$ that must be applied to every prediction in the scenario, or it may provide a different choice of setting values relative to each prediction one wishes to compute. An example of the former would be collapse theories or any interpretation that rejects universal validity of unitary quantum theory, which would require all settings to be 1. An example of the latter is our reasoning rule given by the setting choices illustrated in the completeness result of Theorem C1: In every prediction, e.g., $P(\vec{a}_j = \vec{a}_j|\vec{x} = \vec{\xi})$, the settings for all outcomes that appear in the probability are set to 1, and the settings for the remaining measurements are set to 0.

Now, given an EWFS together with a set of rules $\mathcal{R}$ for choosing the settings in the augmented circuit, we can consider what happens when we design an experiment to observe the outcome referred to in a prediction. Let us do so with a simple example, referring back to Wigner’s original thought experiment, where Alice was the agent and Bob the superagent. When considering the probability of Bob’s outcome b , we have a choice for Alice’s setting x_A and can compute, for instance, $P(b|x_A = 0)$. If it were possible to physically perform a Wigner’s-friend-type experiment involving these agents, and if indeed unitary quantum theory were universally valid, then Bob’s observations regarding b would be consistent with this prediction in the sense of Definition B3: If these premises are satisfied, then Bob can observe $b = \ell$ only if $P(b = \ell|x_A = 0) > 0$. On the other hand, if unitary quantum theory was not universally valid, but there was an additional (yet undiscovered) physical mechanism for objective collapse, then Bob’s physical observations would be consistent with $P(b = \ell|x_A = 1)$ (according to Definition B3) and need not be globally consistent relative to $P(b = \ell|x_A = 0) > 0$. Note that, in these discussions, we have omitted x_B , which

will by default be $x_B = 1$ here since these predictions refer to a nontrivial outcome of Bob.

We have proven in our framework that (setting-conditioned) predictions obtained under any possible rule $\mathcal{R}$ for choosing the settings are mutually consistent. Moreover, as the above example illustrates, when considering observational statements, it is important to consider the physical dynamics leading to the said observations.

From the consistency of the predictions and the definition of global consistency, it is straightforward to see that if the physical dynamics leading to the observations respects the channel choices specified by the reasoning rules $\mathcal{R}$ (and assuming that the physical probabilities also respect the Born rule relative to those channel choices), then the predictions and observations will be globally consistent. On the other hand, if the physical dynamics leading to the observations differs from the $\mathcal{R}$ used to compute the predictions (e.g., when there is physical collapse for all measurements but the rules model certain measurements as pure unitaries), then it is possible to violate global consistency. This provides a way to operationally falsify the rule $\mathcal{R}$ (see also Sec. VII A), assuming that the experiment was performed in a faithful way, in the sense that indeed Bob's operation acted on Alice's whole lab and there was no unexpected information leakage which is not accounted for in the scenario description.

Considering conventional predictions used in FR- and LF-type arguments, we have seen that these imply a particular default rule $\mathcal{R}^{\text{def}}$ for selecting the settings Theorem C1 (as also discussed in Sec. VII A). In this rule, only the measurements whose outcomes appear in a given prediction are assigned projectors (in order to compute said probability through the Born rule), while all other measurements are modeled as pure unitary evolutions. Therefore, if unitary quantum theory were indeed valid universally, our consistency result of Theorem C1 (which refers to only predictive statements) is sufficient to guarantee global consistency of predictive statements obtained through this rule for a given EWFS together with any observations made within a hypothetical experimental realization of that EWFS. Therefore, we have no loss of generality in restricting only to predictive statements in the main text.

1. Consistency and agents' knowledge

This distinction between predictive and observational statements emphasized here provides a precise understanding of what does not constitute an inconsistency. The consistency for predictive statements requires that one should not be able to obtain two different probability assignments P and P' for the same outcomes in a scenario, conditioned on the same information. This refers to only predictions, i.e., probabilities associated with running the protocol for several rounds.

It is important to note that even in a consistent theory (such as purely classical physics) agents can nevertheless assign different probabilities to events due to different knowledge based on observations made in a given round. If Alice and Bob have a fair coin at hand, where Alice knows that the outcome of the coin flip is $c = \text{heads}$ in one round, she would assign probability 1 to $c = \text{heads}$ in that round while Bob would assign a uniform probability to $c = \text{heads}$ if he does not know the outcome. This is not a contradiction, because Alice's probability relates to an observation in a particular round; her certainty in this case would be captured by an observational statement $c = \text{heads}|_{\text{obs}}$. This example leads to consistent predictions, since both agents would predict a uniform probability for $c = \text{heads}$ if asked what the probability of heads will be over many coin flips.

Furthermore, Alice may wish to make a prediction about the outcome b of a bet given that she observed $c = \text{heads}$ (and possibly other information $k = \mathcal{K}$ that she may know about the scenario). This would then correspond to a prediction $P(b = \text{win} | c = \text{heads}, k = \mathcal{K})$. On the other hand, Bob, who does not know the outcome of the coin flip but has the same background information $k = \mathcal{K}$ about the scenario, would make a prediction $P(b = \text{win} | k = \mathcal{K})$ which can be different from $P(b = \text{win} | c = \text{heads}, k = \mathcal{K})$. This is also not a contradiction, since the two predictions are conditioned on different knowledge, and this conditioning is important to ensure consistency. It is immediate to see that, even in these simple classical examples, ignoring the conditioning on agents' knowledge and/or the background assumptions they make about the scenario at hand, one can obtain apparent inconsistencies quite easily.

APPENDIX C: DETAILS OF FORMAL DEFINITIONS AND KEY PROPERTIES OF OUR FRAMEWORK

In this section, we provide full formal definitions of certain concepts (such as causal and nonsuperagent structure of an EWFS and the assumptions **Q**, **U**, **C**, **D**, and **S**) that were motivated intuitively in Secs. IV A, IV B, and VI in the main text and thereby formally present a theorem which proves completeness, consistency, and causality properties of our framework.

1. Completeness, consistency, and causality

As explained in Sec. IV A, even though our augmented circuit formalism does not *a priori* impose absoluteness of events, it can be shown that it provides a complete representation of all predictions that can be made in an EWFS in a way that is consistent and respects causality principles. Before stating the formal theorem to establish this, we define what we mean by causality principles. For this, we begin by defining a directed acyclic graph (DAG) [66] that corresponds to every EWFS. This DAG captures

the potential information flow within the EWFS protocol, adhering to the time-ordered sequence of operations and representing an objective and agent independent property of that EWFS. We refer to this as the causal structure of the EWFS.

Recall that, by definition, an EWFS with N agents is modeled as having $2N$ operations without loss of generality, as each agent performs a measurement $\mathcal{M}^{A_i}$ followed by some quantum channel $\mathcal{E}_i$, both associated with a time step t_i . Let $\mathcal{O}$ denote any one of these $2N$ operations; we denote by $S_{\mathcal{O}}$ the subset of all systems and memories $S \cup M$ such that the operation $\mathcal{O}$ acts nontrivially on all systems in $S_{\mathcal{O}}$ and as the identity on the rest of $S \cup M$.

Definition C1 (causal structure of an EWFS). The causal structure of an EWFS with N agents is a DAG G with the following properties:

- (1) G has $2N$ vertices given by the set $\text{Vert}(G) := \{V_i^M, V_i^E\}_{i=1}^N$, where V_i^M and V_i^E are, respectively, associated with the operations $\mathcal{M}^{A_i}$ and $\mathcal{E}_i$, and both pairs of such vertices are associated with the time step t_i , for each i .
- (2) G contains a directed edge $V \rightarrow V'$ for $V, V' \in \text{Vert}(G)$ whenever $t_V < t_{V'}$ and $S_{\mathcal{O}_V} \cap S_{\mathcal{O}_{V'}} \neq \emptyset$, where $\mathcal{O}_V$ and $\mathcal{O}_{V'}$ are the operations and t_V and $t_{V'}$ are the time steps associated with the vertices V and V' , respectively.

Definition C2 (directed paths and partial order). The DAG G associated with an EWFS in Definition C1 defines a partial order relation $<$ on agents in the EWFS, with $A_i < A_j$ if and only if there is a directed path from the measurement vertex V_i^M of A_i to the measurement vertex V_j^M of A_j . Furthermore, we write $A_i <^S A_j$ whenever there is such a directed path and the system S is included in at least one of the set intersections $S_{\mathcal{O}_V} \cap S_{\mathcal{O}_{V'}}$ involved in that path. We use $A_i \not< A_j$ and $A_i \not<^S A_j$ to denote the absence of directed paths with the above-defined properties.

Crucially, notice that the above graph G and induced partial order $<$ represent objective properties inherent to the EWFS, independent of the specific settings chosen to model the measurement $\mathcal{M}^{A_i}$ in the augmented EWFS. This is because the definition relies on only the time order of operations and the set of systems on which the operations act. These aspects are included in the description of the original EWFS, Definition 1, and the settings do not affect them.

Given this definition, a natural causality principle is that an outcome a_j should not depend on a setting ξ_i of a measurement whenever $A_i \not< A_j$. As our circuits are acyclic and operations therein have a clear time ordering, we have $A_i < A_j$ implies $t_i < t_j$, and it is easy to see that this ensures that G is indeed a directed acyclic graph and that $<$ is a partial order relation. Moreover, this causality principle ensures that there is no retrocausal dependence of outcomes on future settings, despite settings in our formalism

representing Heisenberg cuts rather than actual experimental measurement choices (further elaborated in Sec. VII C).

Finally, we note that $t_i < t_j$ does not necessarily imply $A_i \not< A_j$ as it is possible to have two measurements acting at different times but on nonoverlapping sets of systems; then, G will not contain any directed paths between the measurements of these agents. Thinking of different subsystems as being embedded at different “spatial locations,” we can regard such disjoint sets of systems as being spacelike separated. Therefore, if we consider the circuit as being embedded in a spacetime, with an output of one operation connected to the input of another only if the former is in the past light cone of the latter, then $<$ is compatible with the causal structure of the spacetime [60,61].

Theorem C1.

- (1) *Completeness*—In any given EWFS, all conventional predictions in that EWFS can be derived within the single augmented circuit of that EWFS. More explicitly, each conventional prediction $P_{\text{conv}}(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l)$ in the EWFS equals a particular setting conditioned prediction $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi}^*)$ of the augmented circuit where the setting choice $\vec{x} = \vec{\xi}^*$ is such that $x_i = 1$ for all $i \in \{j_1, \dots, j_p, l_1, \dots, l_q\}$ and $x_i = 0$ for all $i \notin \{j_1, \dots, j_p, l_1, \dots, l_q\}$.
- (2) *Consistency*—For any EWFS, the set of all statements Σ^{aug} obtained in the corresponding augmented circuit (Definition 10) are consistent according to Definition 5.
- (3) *Causality*—For every setting-conditioned prediction $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi})$ and every i such that $A_i \not< A_k$ for all $k \in \{j_1, \dots, j_p, l_1, \dots, l_q\}$, the prediction is independent of the setting x_i . That is, for all such i , we have the following, noting that $\vec{x} = \vec{\xi}$ is short for $x_1 = \xi_1, \dots, x_N = \xi_N$ and each $\xi_i \in \{0, 1\}$:

$$\begin{aligned} P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, (x_1 = \xi_1, \dots, x_i = 0, \dots, x_N = \xi_N)) \\ = P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, (x_1 = \xi_1, \dots, x_i = 1, \dots, x_N = \xi_N)). \end{aligned} \quad (\text{C1})$$

Whenever a setting-conditioned prediction is independent of a setting x_i , i.e., satisfies Eq. (J10), we simply drop x_i from that prediction and denote it as follows:

$$\begin{aligned} P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi}) \\ = P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} \setminus x_i = \vec{\xi} \setminus \xi_i), \end{aligned} \quad (\text{C2})$$

where $\vec{x} \setminus x_i = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_N)$ is the setting vector with the i th component omitted, and similarly for the vector of setting values $\vec{\xi} \setminus \xi_i$.

A proof of this theorem is deferred to Appendix J 2, as it uses results from Appendix E.

2. Properties of agents' reasoning

In Sec. IV B, we motivated the assumptions **Q**, **U**, **C**, **D**, and **S** which formalize a version of FR's assumptions Q , U , C , D , and S while making explicit the role of the settings or Heisenberg cuts. Our main result concerning agents' reasoning in EWFS was to show that this formal version of assumptions regarding agents' use of unitary quantum theory to each others' labs and classical logic applied to observed outcomes is, in fact, consistent in any EWFS. Here, we provide the formal definitions of these assumptions which utilize methods from epistemic modal logic for defining C , D , and S (inspired by previous work [4,28]).

One of the assumptions (or logical axioms) used in the FR argument is of the form “If Alice is certain that Bob is certain that the outcome $a = 1$,” then “Alice is certain that $a = 1$.” This is called the C assumption there and is about the ability of Alice to inherit Bob's knowledge. In Ref. [4], this assumption C was formalized within the framework of modal logic. Here, $K_{A_i}(S)$ is used to denote that “agent A_i knows that the statement S is true,” where K^{A_i} is known as a knowledge operator (see Ref. [4] for a formal definition of the knowledge operators in terms of Kripke structures in modal logic). The assumption C can then be succinctly expressed as an inference of the form

$$K_{A_i}K_{A_j}(S) \Rightarrow K_{A_i}(S). \quad (\text{C3})$$

More specifically, it is pointed out in Ref. [4] that such an inference needs to be made only when the agent A_i *trusts* the agent A_j (for otherwise A_i may not believe in everything that A_j claims to know and may not want to inherit A_j 's knowledge). In Wigner's friend scenarios such as that of FR, Ref. [4] instantiates the trust structure by considering pairs of agents performing compatible measurements and applies C only for such pairs. The trust structure will not be relevant for the general solution that we propose here because, as we show (cf. Theorem C1), the inclusion of the settings in our framework ensures the general validity of Eq. (C3) in an augmented EWFS independently of the trust structure.

Definition C3 (Assumption C). For any two agents A_i and A_j , Eq. (C3) holds for all statements $S \in \Sigma^{\text{aug}}$.

As noted in Ref. [4], the FR argument also uses the distributive axiom of logic. We instantiate this in our formalism below.

Definition C4 (Assumption D). For any set of agents reasoning using the augmented circuit, if $S_1 \in \Sigma_L^{\text{aug}}$ and $S_1 \Rightarrow S_2$, then $S_2 \in \Sigma_L^{\text{aug}}$ holds and we have

$$K_{A_i}(S_1 \wedge (S_1 \Rightarrow S_2)) \Rightarrow K_{A_i}(S_2); \quad (\text{C4})$$

that is, if an agent A_i knows S_1 and also that S_1 implies S_2 , then the agent knows S_2 .

Finally, we consider the S assumption of FR which states that an agent cannot be certain of two opposite values—say, $a = 0$ and $a = 1$ —of a measurement outcome. In our framework, this is a weaker version of our general consistency condition of Definition 5, as the latter applies to all predictions while the former only to logical predictions.

Definition C5 (Assumption S). For any subset $\vec{a}_j$ of measurement outcomes, if $P(\vec{a}_j = \vec{a}_j) = 1$, then it is impossible to have $P(\vec{a}_j = \vec{a}'_j) = 1$ for any outcome values $\vec{a}_j \neq \vec{a}'_j$.

Now, we present the quantum theory-dependent assumptions which independently capture the validity of the Born rule and of unitary evolution of closed quantum systems (including agents' labs).

Definition C6 (Assumption Q). Consider a statement $S :=$ “If the outcomes $\vec{a}_l$ take values $\vec{a}_l$ and the settings take the value $\vec{x} = \xi$, then the outcomes $\vec{a}_l$ take values $\vec{a}_l$ with a probability P .” Agents in an EWFS can regard such a statement S as true if and only if the corresponding setting-conditioned prediction $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \xi) = P$ can be derived by applying the Born rule to the EWFS (as detailed in Definition 8).

We note that, by construction, the set of all statements Σ^{aug} constructed from setting-conditioned predictions in the augmented circuit has this property. In our formalization of the **U** assumption, we make explicit another implicit assumption in the previous literature, which relates to quantum control over other agents' labs.

Definition C7 (Assumption U). Agents can choose the setting $x_i = 0$ (i.e., pure unitary description) for the measurement of any agent A_i whose outcome they are not logically reasoning about; i.e., for every setting-conditioned prediction $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \xi)$, both choices $x_i \in \{0, 1\}$ are allowed for all $i \notin \{j_1, \dots, j_p, l_1, \dots, l_q\}$. Moreover, agents can have full quantum control over the labs of other agents; i.e., the measurement of an agent A_i can act nontrivially on the total system $S_j M_j$ comprising the lab of another agent A_j .

3. Nonsuperagent structure

The concept of the nonsuperagent structure nSA of an EWFS was explained and illustrated in Sec. VI, and it defines a criterion for when one agent does not act as a superagent to another agent in an EWFS, and thereby distinguishes standard quantum experiments from genuinely Wigner's-friend-like experiments (Definition 13). This criterion, defined below, is based on the causal structure and partial order or DAG of an EWFS defined in Definitions C1 and C2 presented earlier in this appendix.

Definition C8 (nonsuperagent structure nSA). We say that an EWFS involving a set $\{A_1, \dots, A_N\}$ of agents respects a nonsuperagent structure nSA whose elements are pairs (A_i, A_j) of agents, if the given EWFS is operationally equivalent to another EWFS involving a

set $\{A'_1, \dots, A'_N\}$ of agents, such that for all $(A_i, A_j) \in n\mathcal{SA}$ the following conditions hold.

- (1) If $i = j$, then $\mathcal{E}'_i$ acts trivially on the memory M'_i .
- (2) If $i \neq j$, then one of the following holds:
 - (i) $A'_i \not\prec A'_j$.
 - (ii) If $A'_i \prec A'_j$, then $\mathcal{E}'_i$ acts trivially on the memory M'_i and $A'_i \not\prec^{M'_i} A'_j$.

If $(A_i, A_j) \in n\mathcal{SA}$, we say that A_j does not act on the memory of A_i or does not act as a superagent to A_i .

APPENDIX D: OVERVIEW OF THE FRAUCHIGER-RENNER APPARENT PARADOX

1. The FR no-go theorem

Here, we review the assumptions Q , C , and S of FR's claimed no-go theorem, as well as the additional assumptions U and D that are also relevant to the FR analysis, as pointed out in Ref. [4]. The FR protocol and formal statement of their no-go theorem are reviewed in Appendix D 3.

Assumption (Q)—It asserts that an agent can be certain that a given proposition holds whenever the quantum-mechanical Born rule assigns probability 1 to it. Specifically, suppose that agent A has established Statement A⁽ⁱ⁾: “System S is in state $|\psi\rangle_S$ at time t_0 .” Suppose furthermore that agent A knows Statement A⁽ⁱⁱ⁾: “The value x is obtained by a measurement of S with respect to the family $\{\pi_x^{t_0}\}_{x \in \mathcal{X}}$ of Heisenberg operators relative to time t_0 , which is completed at time t .” If $\langle \psi | \pi_\xi^{t_0} | \psi \rangle = 1$ for some $\xi \in \mathcal{X}$, then agent A can conclude Statement A⁽ⁱⁱⁱ⁾: “I am certain that $x = \xi$ at time t .”

Assumption (C)—It asserts that one agent can inherit the knowledge of another agent who uses the same theory as them to arrive at their conclusions. Suppose that agent A has established Statement A⁽ⁱ⁾: “I am certain that agent A, upon reasoning within the same theory as the one I am using, is certain that $x = \xi$ at time t .” Then agent A can conclude Statement A⁽ⁱⁱ⁾: “I am certain that $x = \xi$ at time t .”

Assumption (S)—From the viewpoint of an agent who carries out a particular measurement, this measurement has one single outcome. Specifically, suppose that agent A has established Statement A⁽ⁱ⁾: “I am certain that $x = \xi$ at time t .” The agent A must necessarily deny Statement A⁽ⁱⁱ⁾: “I am certain that $x \neq \xi$ at time t .”

Note that a violation of S can itself be interpreted as a logical paradox, as it would imply that the outcome x is both ξ and not ξ with certainty.

In Ref. [4], the authors also identify additional assumption U (unitarity) that FR use (as part of Q) but did not explicitly state in their set of assumptions.

Assumption (U)—An agent A can model measurements performed by any other agent B as reversible unitary evolutions in B's lab.

Furthermore, they also note that the FR reasoning involves another basic rule of classical logical inference, namely, the *distributive axiom*.

Assumption (D)—If an agent A knows a statement s_1 and also knows that s_1 implies another statement s_2 , then agent A can conclude that they know s_2 .

They formulate this in terms of a knowledge operator which formally keeps track of which agent knows which statement [4]. We also introduce said notation later. Assumption D is implicitly used in FR when the statements s_1 and s_2 correspond to measurement outcomes; e.g., s_1 could be “I am certain that $x = \xi$ at time t ,” s_2 could be “I am certain that $x' = \xi'$ at time t' .” Then if agent A who knows s_1 also knows “If I am certain that $x = \xi$ at time t , then I am certain that $x' = \xi'$ at time t' ,” they would use D to conclude that agent A is certain of the statement s_2 , “I know that $x' = \xi'$ at time t' .”

In Ref. [4], the FR argument is refined by making explicit the additional assumptions U and D . Thus, they suggest the implication that the assumptions Q , U , C , D , and S cannot be simultaneously satisfied in the protocol proposed by FR. We note that this result is proven in Ref. [4] by formalizing these assumptions using the Kripke structure of epistemic modal logic (which is the branch of logic that refers to knowledge of agents), and we refer the reader to Ref. [4] for the formal statement of this theorem in this mathematical language. We do not review the full modal logic framework here but refer to aspects of it wherever necessary.

At a broad level, the proof proceeds by considering the FR scenario where agents reason about each other's knowledge using assumptions Q , U , C , and D and claims to show that such agents would always arrive at a violation of S , which as we have explained above can be interpreted as a paradox.

2. Entanglement version of the FR scenario

In Sec. V and Fig. 4 in the main text, we provided an overview of the entanglement version of the FR scenario and the main arguments. Here, we review in more detail the entanglement-based version of the FR scenario and apparent paradox that highlights the proof method typically employed to prove the no-go claim regarding a contradiction between the assumptions Q , U , C , and D (reviewed in the subsection above).

The entanglement version of the FR protocol is originally attributed to Luis Masanes (based on a talk) and was also mentioned by Matthew Pusey in Ref. [10]. It is much simpler than FR's original (prepare-and-measure-based) protocol but makes the important and salient features of the FR protocol more readily accessible. The resolution to EWFS paradoxes given by our work is, however, fully general and applies, in particular, to both the entanglement and the original prepare-and-measure version of FR's arguments. For the interested reader, we provide a review

of the original FR scenario in Appendix D 3 and apply our framework to resolve it, in Appendix G 2.

In this section, we follow the notation and agent naming conventions of Refs. [4,28]. Here, we have two agents Alice and Bob, who measure individual subsystems of a bipartite system, while two superagents Ursula and Wigner measure the labs (system and memory) of Alice and Bob, respectively. The protocol can be broken down into three steps: a bipartite state preparation (preselection) at an initial time $t = 1$, intermediate local measurements by Alice and Bob on the state at time $t = 2$, and a final measurement and postselection by two superagents Ursula and Wigner at time $t = 3$.

- (i) *Preselection at time $t = 1$* —An initial state $|\psi^{t=1}\rangle_{RS} := (1/\sqrt{3})(|00\rangle_{RS} + |10\rangle_{RS} + |11\rangle_{RS})$ is prepared and shared between Alice and Bob, where R and S label the subsystems belonging to Alice and Bob, respectively. Alice and Bob's memories A and B are initialized to $|0\rangle_A$ and $|0\rangle_B$, respectively. Hence, the initial preparation (i.e., preselected state) on RASB is given by

$$|\psi^{t=1}\rangle_{RASB} := \frac{1}{\sqrt{3}}(|0000\rangle + |1000\rangle + |1010\rangle)_{RASB}. \quad (D1)$$

- (ii) *Intermediate operations at time $t = 2$* —Alice and Bob measure their respective systems in the computational basis $\{|0\rangle, |1\rangle\}$ and store the outcomes a and b of their respective measurements in their memory systems. From the outside perspective, Alice's measurement is modeled as a unitary $\mathcal{M}_{\text{unitary}}^A := (\text{CNOT})_{RA}$ on the joint system RA which performs a CNOT operation with R as control and A as target, while Bob's measurement is similarly modeled as a unitary $\mathcal{M}_{\text{unitary}}^B := (\text{CNOT})_{SB}$.
- (iii) *Postselection at time $t = 3$* —The superagents Ursula and Wigner postselect on the following final state: $|\phi^{t=3}\rangle_{RASB} := |\text{ok}\rangle_{RA}|\text{ok}\rangle_{SB}$, which they achieve by measuring RA and SB, respectively, in the basis $\{|\text{ok}\rangle := (1/\sqrt{2})(|00\rangle - |11\rangle), |\text{fail}\rangle := (1/\sqrt{2})(|00\rangle + |11\rangle)\}$ and halting when both of them obtain the outcome $u = \text{ok}$, $w = \text{ok}$ (where u and w denote Ursula's and Wigner's outcomes, respectively, and the halting condition is checked by announcing these measurement outcomes in each run).

When modeling both measurements as unitaries, the joint state of RASB just after time $t = 2$ is given by

$$\begin{aligned} |\psi^{t=2}\rangle_{RASB} &= (\mathcal{M}_{\text{unitary}}^A \otimes \mathcal{M}_{\text{unitary}}^B) |\psi^{t=1}\rangle_{RASB} \\ &= \frac{1}{\sqrt{3}}(|0000\rangle + |1100\rangle + |1111\rangle)_{RASB}. \end{aligned} \quad (D2)$$

The superagents can then calculate the probability of success of the postselection given the preselection and intermediate unitary evolution as

$$\begin{aligned} P(u = w = \text{ok} | \psi^{t=2}) &= |\langle \phi^{t=3} | \psi^{t=2} \rangle|^2 \\ &= |\langle \phi^{t=3} | (\mathcal{M}_{\text{unitary}}^A \otimes \mathcal{M}_{\text{unitary}}^B) | \psi^{t=1} \rangle|^2 \\ &= \frac{1}{12}. \end{aligned} \quad (D3)$$

By this, they establish that they have a nonzero probability of obtaining $u = w = \text{ok}$ and can, thus, repeat the protocol until they succeed. Upon successfully obtaining the desired outcomes, the protocol is halted and the agents reason about each others' knowledge as follows, where we recall that $K_A(S)$ denotes that agent A knows the statement S .

- (i) Upon obtaining $u = \text{ok}$ on measuring RA, Ursula reasons using the joint state $|\psi^{t=2}\rangle$ [Eq. (D2)] that Bob must have certainly obtained the outcome $b = 1$ upon measuring S, since $\langle \text{ok} |_{RA} \langle 00 |_{SB} | \psi^{t=2} \rangle_{RASB} = 0$. This gives

$$K_U(u = \text{ok} \Rightarrow b = 1). \quad (D4)$$

- (ii) Using the same state, Ursula knows that if Bob obtained $b = 1$ on measuring S, he would have concluded with certainty that Alice obtained $a = 1$ on measuring R, since $\langle 00 |_{RA} \langle 11 |_{SB} | \psi^{t=2} \rangle_{RASB} = 0$:

$$K_U K_B(b = 1 \Rightarrow a = 1). \quad (D5)$$

- (iii) Again using the same state, Ursula further reasons that Bob knows that if Alice had obtained $a = 1$, she would have concluded with certainty that Wigner would obtain $w = \text{fail}$, since $\langle 11 |_{RA} \langle \text{ok} |_{SB} | \psi^{t=2} \rangle_{RASB} = 0$. This gives

$$K_U K_B K_A(a = 1 \Rightarrow w = \text{fail}). \quad (D6)$$

As shown in Ref. [4], the above three statements can be combined using the assumption C of the FR paper [in the form of Eq. (C3)] and the distributive axiom D to yield the following paradoxical chain of statements. To obtain this result, the assumption C in the form of Eq. (C3) needs only to be used between the following pairs of agents: Alice and Bob, Alice and Wigner, Ursula and Bob, and Ursula and Wigner, as other pairs of agents need not trust each other [4,28,67]:

$$K_U(u = \text{ok} \Rightarrow b = 1 \Rightarrow a = 1 \Rightarrow w = \text{fail}), \quad (D7)$$

or, in short, $K_U(u = \text{ok} \Rightarrow w = \text{fail})$. This argument aims to establish that agents reasoning using Q , U , C , and D will arrive at a contradiction with S , as they conclude through such a reasoning that $w = \text{ok}$ and $w = \text{fail}$ must both hold with certainty and is, therefore, regarded as a proof of the FR no-go theorem (and its refinement as given

in Ref. [4]) regarding the incompatibility of Q , U , C , D , and S .

3. Prepare-and-measure version of the FR scenario

The scenario in Ref. [3] describes a protocol realized by agents F , $\bar{F}$, W , and $\bar{W}$. Agents F and $\bar{F}$ have their own individual labs, while W and $\bar{W}$ are so-called superagents; i.e., W can perform arbitrary measurements on F and the lab of F , while $\bar{W}$ can perform arbitrary measurements on $\bar{F}$ and the lab of $\bar{F}$. The labs of F and $\bar{F}$ are completely isolated from W and $\bar{W}$ until W and $\bar{W}$ measure at the end of the protocol. It is also assumed that the labs F and $\bar{F}$ are initially in pure states. The lab systems of F and $\bar{F}$ are denoted L and $\bar{L}$, respectively. $\bar{L}$ includes everything in $\bar{F}$'s lab such as the agent $\bar{F}$ and a random generator R that they use but excludes a spin qubit S which will start off in the lab of $\bar{F}$ and move to the lab of F during the protocol. The lab system L of F will include the spin S that arrives to F as well as the agent F and other devices in their lab which are not explicitly specified.

The protocol is repeated n times. Between each implementation, it is reset to the initial state. There is a halting condition which is examined at the end of each round. When the condition is satisfied, the protocol is stopped and the last round of the experiment is analyzed. The n th round of the protocol from the perspective of W and $\bar{W}$ is as follows.

Before time $n:00$, $\bar{F}$ tosses a coin in her lab which gives heads with probability $1/3$ and tails with probability $2/3$. This coin toss is a random variable which is obtained by measuring the following quantum state in the same basis in which it is expressed:

$$|\text{init}\rangle_R := \sqrt{\frac{1}{3}}|\text{heads}\rangle_R + \sqrt{\frac{2}{3}}|\text{tails}\rangle_R. \quad (\text{D8})$$

The lab of $\bar{F}$ consists of the random generator (or coin) R , a spin S , and $\bar{L}\setminus$, which represents the rest of the lab. In the following, $\bar{L} := R \otimes \bar{L}\setminus$ is the lab of $\bar{F}$, excluding the spin S . Then, the unitary $U_{\bar{F}}$ (from the perspective of F , W , and $\bar{W}$) that describes $\bar{F}$'s coin toss (measurement of R) implements the evolution

$$U_{\bar{F}}: |\text{init}, \phi_0, S_0\rangle_{\bar{L}S} \rightarrow \left(\sqrt{\frac{1}{3}}|\bar{h}\rangle_{\bar{L}} + \sqrt{\frac{2}{3}}|\bar{t}\rangle_{\bar{L}} \right) |S_0\rangle_S, \quad (\text{D9})$$

where $|\text{init}, \phi_0, S_0\rangle_{\bar{L}S} := |\text{init}\rangle_R |\phi_0\rangle_{\bar{L}\setminus} |S_0\rangle_S$ with $|\phi_0\rangle_{\bar{L}\setminus}$ the initial state of $\bar{F}$'s lab excluding the spin system and the random generator—the latter two being $|S_0\rangle_S$ and $|\text{init}\rangle_R$, respectively, where $|S_0\rangle_S$ is some initial state of S . The kets $|\bar{h}\rangle$ and $|\bar{t}\rangle$ represent the state of $\bar{F}$'s lab (excluding the spin qubit) after her measurement of the random variable ($|\bar{h}\rangle$ in

the case her measurement revealed heads while $|\bar{t}\rangle$ when she obtained tails).

Following the coin toss, $\bar{F}$ prepares a spin qubit in her lab in state $|\downarrow\rangle_S$ if she gets heads while $|\rightarrow\rangle_S$ if she gets tails. From the perspective of F , W , and $\bar{W}$, this is a unitary process, since her lab is an isolated system. The unitary that describes this preparation of S is denoted as U_{spin} , and, together with $U_{\bar{F}}$, this implements the evolution

$$U_{\text{spin}} U_{\bar{F}}: |\text{init}, \phi_0, S_0\rangle_{\bar{L}S} \rightarrow \sqrt{\frac{1}{3}}|\bar{h}, \downarrow\rangle_{\bar{L}S} + \sqrt{\frac{2}{3}}|\bar{t}, \rightarrow\rangle_{\bar{L}S}. \quad (\text{D10})$$

Meanwhile, F waits patiently in her lab, which we denote $|\phi_0\rangle_{L\setminus}$ initially, where $L\setminus$ denotes F 's lab without the spin qubit S and we use $L := S \otimes L\setminus$. Between times $n:10$ and $n:20$, $\bar{F}$ sends her qubit system to F 's lab, and F subsequently measures the qubit in the $|\uparrow\rangle_S, |\downarrow\rangle_S$ basis, denoting her outcome $+1/2$ and $-1/2$, respectively. From the perspective of W and $\bar{W}$, the labs of F and $\bar{F}$ are subsequently modeled through the series of evolutions $U_{\bar{F}}$, followed by U_{spin} followed by U_F , where U_F is the unitary evolution corresponding to F 's measurement of S :

$$\begin{aligned} U_F U_{\text{spin}} U_{\bar{F}}: |\text{init}, \phi_0, S_0\rangle_{\bar{L}S} |\phi_0\rangle_{L\setminus} \\ \rightarrow \sqrt{\frac{1}{3}} \left(|\bar{h}\rangle_{\bar{L}} |\downarrow, -1/2\rangle_L + |\bar{t}\rangle_{\bar{L}} |\downarrow, \right. \\ \left. -1/2\rangle_L + |\bar{t}\rangle_{\bar{L}} |\uparrow, 1/2\rangle_L \right), \end{aligned} \quad (\text{D11})$$

where $|\downarrow, -1/2\rangle_L$ denotes the spin in the down state $|\downarrow\rangle_S$ and the state of the rest of lab L when F obtains measurement outcome $-1/2$; similarly for $|\downarrow, -1/2\rangle_L$. Between times $n:20$ and $n:30$, $\bar{W}$ measures lab $\bar{L}$ in the basis $\{|\overline{\text{ok}}\rangle_{\bar{L}}, |\overline{\text{fail}}\rangle_{\bar{L}}\}$, where $|\overline{\text{ok}}\rangle_{\bar{L}} := \sqrt{\frac{1}{2}}(|\bar{h}\rangle_{\bar{L}} + |\bar{t}\rangle_{\bar{L}})$, $|\overline{\text{fail}}\rangle_{\bar{L}} := \sqrt{\frac{1}{2}}(|\bar{h}\rangle_{\bar{L}} - |\bar{t}\rangle_{\bar{L}})$. After time $n = 30$, W measures lab L in the basis $\{|\text{ok}\rangle_L, |\text{fail}\rangle_L\}$, where $|\text{ok}\rangle_L := \sqrt{\frac{1}{2}}(|\downarrow, -1/2\rangle_L - |\uparrow, 1/2\rangle_L)$ and $|\text{fail}\rangle_L := \sqrt{\frac{1}{2}}(|\downarrow, -1/2\rangle_L + |\uparrow, 1/2\rangle_L)$. The final state of the labs [right-hand side of Eq. (D11)] in this measurement basis takes on the form

$$\begin{aligned} \sqrt{\frac{1}{12}}(|\overline{\text{fail}}\rangle_{\bar{L}} + |\overline{\text{ok}}\rangle_{\bar{L}})(|\text{fail}\rangle_L + |\text{ok}\rangle_L) \\ + \sqrt{\frac{1}{6}}(|\overline{\text{fail}}\rangle_{\bar{L}} - |\overline{\text{ok}}\rangle_{\bar{L}})|\text{fail}\rangle_L. \end{aligned} \quad (\text{D12})$$

In Fig. 6, we summarize this scenario using branching notation. If the outcomes of W and $\bar{W}$ are not ok and $\overline{\text{ok}}$, respectively, then the protocol is reset and repeated. If their outcomes are ok and $\overline{\text{ok}}$, then the experiment is finished and

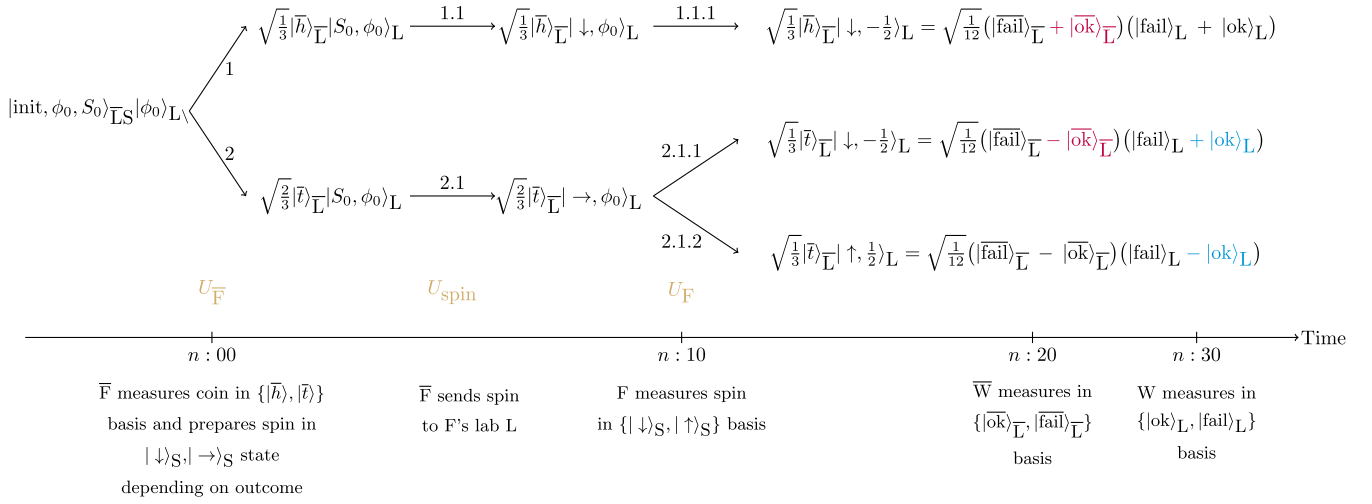


FIG. 6. The FR scenario expressed using branching notation. The quantum state, as per a unitary description from the viewpoint of W and $\bar{W}$ at any time, is the sum of the states of each branched at said time. Arrows indicate the splitting of the amplitudes due to the actions of individual observers. We refer to these as *branches*, and each branch has its own number above the corresponding arrow. The two purple kets on branches 1.1.1 and 2.1.1 cancel each other out when said branches are added together, while the two blue kets on branches 2.1.1 and 2.1.2 cancel when said branches are added together.

F , $\bar{F}$, W , and $\bar{W}$ reason about their ok and $\bar{\text{ok}}$ measurement outcomes and what they should be able to predict about the measurement outcomes of the other agents in the last round of the protocol. The obtention of ok and $\bar{\text{ok}}$ by W and $\bar{W}$, respectively, is the halting condition mentioned previously, and, from the above form of the final state, it is clear that this outcome occurs with probability $\frac{1}{12}$.

While reasoning, they assume that all agents are aware of the entire experimental procedure as described above and that they all employ the same theory. As we see, all agents, when reasoning from *their* perspective, assume unitary dynamics for all agent's measurements other than those they are reasoning about. In other words, their protocol is a special case of an EWFS. In Fig. 9, we present the augmented circuit of the original FR protocol as a simplified version that makes the mapping back to the entanglement version more apparent.

In FR's reasoning, they make the three assumptions Q , C , and S explicitly, in addition to the assumptions U and D implicitly, as described in Appendix D 1.

In particular, the authors then claim to prove the following theorem.

Theorem D1. Any theory that satisfies assumptions (Q), (C), and (S) yields contradictory statements when applied

to their thought experiment of box 1. By “box 1,” the authors are referring to the above protocol.

APPENDIX E: DERIVATION OF PREDICTIONS IN THE AUGMENTED EWFS

This section serves the purpose of deriving explicit expressions for predictions and setting-conditioned predictions in the augmented circuit of an EWFS from the Born rule and probability theory.

Our sample space Ω is chosen to be the set of all measurement outcomes under all settings, namely,

$$\Omega = \{a_1, x_1, a_2, x_2, \dots, a_N, x_N\}_{\{a_j, x_j\}_j}, \quad (\text{E1})$$

where $a_j \in \perp$ for $x_j = 0$ and $a_j \in \text{O}_j$ for $x_j = 1$. The set of events is the power set of Ω . It follows from applying the Born rule to our augmented circuit (see also Fig. 3) that if agents $A_1, A_2, \dots, A_N$ were to perform measurements under settings $\vec{x}$ with outcomes corresponding to projectors $\pi_{a_1, x_1}^{A_1}, \pi_{a_2, x_2}^{A_2}, \dots, \pi_{a_N, x_N}^{A_N}$, respectively, then the probability of these elementary events is

$$\begin{aligned} P(a_1, a_2, \dots, a_N, \vec{x}) &= P(a_1, a_2, \dots, a_N | \vec{x}) P(\vec{x}) \\ &= \text{tr} \left[\mathcal{E}^N \left(\pi_{a_N, x_N}^{A_N} \mathcal{M}_{\text{unitary}}^{A_N} \left(\dots \mathcal{E}^2 \left(\pi_{a_2, x_2}^{A_2} \mathcal{M}_{\text{unitary}}^{A_2} \left(\mathcal{E}^1 \left(\pi_{a_1, x_1}^{A_1} \mathcal{M}_{\text{unitary}}^{A_1} (\rho_0) \pi_{a_1, x_1}^{A_1} \right) \right) \pi_{a_2, x_2}^{A_2} \right) \dots \right) \pi_{a_N, x_N}^{A_N} \right) \right] P(\vec{x}), \quad (\text{E2}) \end{aligned}$$

where ρ_0 is the initial “preselected” state, namely, $\rho_{S_1, \dots, S_m} \otimes |0\rangle\langle 0|^{\otimes N}$, and $P(\vec{x})$ the unconditional probability distribution over settings. The other term in Eq. (E2) is the probability of outcomes $\{a_1, a_2, \dots, a_N\}$ conditioned on setting $\vec{x}$.

The probability of the other events can be derived by marginalizing over this distribution. In particular, we are interested in the probability of events corresponding to agents $\{A_{j_1}, A_{j_2}, \dots, A_{j_p}\}$ obtaining outcomes $\{a_{j_1}, a_{j_2}, \dots, a_{j_p}\}$, conditioned on the setting being $\vec{x}$. This is given by

$$P(a_{j_1}, a_{j_2}, \dots, a_{j_p} | \vec{x}) = \sum_{\substack{a_1 \text{ if } x_1=1 \& 1 \notin \text{OUT} \\ a_2 \text{ if } x_2=1 \& 2 \notin \text{OUT} \\ \vdots \\ a_N \text{ if } x_N=1 \& N \notin \text{OUT}}} \text{tr} \left[\mathcal{E}^N \left(\pi_{a_N, x_N}^{\mathcal{A}_N} \dots \mathcal{E}^2 \left(\pi_{a_2, x_2}^{\mathcal{A}_2} \mathcal{M}_{\text{unitary}}^{\mathcal{A}_2} \left(\mathcal{E}^1 \left(\pi_{a_1, x_1}^{\mathcal{A}_1} \mathcal{M}_{\text{unitary}}^{\mathcal{A}_1} (\rho_0) \pi_{a_1, x_1}^{\mathcal{A}_1} \right) \right) \pi_{a_2, x_2}^{\mathcal{A}_2} \right) \dots \pi_{a_N, x_N}^{\mathcal{A}_N} \right) \right], \quad (\text{E3})$$

$\text{OUT} := \{j_1, j_2, \dots, j_p\}$. We have not summed over settings $x_j = 0$, since the corresponding observer is modeled unitarily with a deterministic outcome, $a_i = \perp$. Note that if we have a set of projectors $\{\pi_j\}_j$ on a system A and a quantum channel $\mathcal{E}$ also on A, from the Stinespring dilation theorem and other elementary properties, it follows that there exists another system B such that for an arbitrary linear operator $\hat{A}$ on A we have

$$\begin{aligned} \sum_j \text{tr}_A [\mathcal{E}(\pi_j \hat{A} \pi_j)] &= \sum_j \text{tr}_A [\text{tr}_B [U_{AB}(\pi_j \hat{A} \pi_j) \otimes \rho_B U_{AB}^\dagger]] \\ &= \sum_j \text{tr}_{AB} [U_{AB}(\pi_j \hat{A} \otimes \rho_B \pi_j) U_{AB}^\dagger] \\ &= \text{tr}_{AB} \left[\left(\sum_j \pi_j^2 \right) \hat{A} \otimes \rho_B \right] = \text{tr}_A [\hat{A}]. \end{aligned} \quad (\text{E4})$$

Applying this equality iteratively to Eq. (E3) allows us to simplify it:

$$\begin{aligned} P(a_{j_1}, a_{j_2}, \dots, a_{j_p} | \vec{x}) &= \sum_{\substack{a_1 \text{ if } x_1=1 \& 1 \notin \text{OUT} \\ a_2 \text{ if } x_2=1 \& 2 \notin \text{OUT} \\ \vdots \\ a_{K-1} \text{ if } x_{K-1}=1 \& K-1 \notin \text{OUT}}} \text{tr}[\ast\ast], \\ \ast\ast &= \pi_{a_K, 1}^{\mathcal{A}_K} \mathcal{E}^{K-1} \left(\pi_{a_{K-1}, x_{K-1}}^{\mathcal{A}_{K-1}} \dots \mathcal{E}^2 \left(\pi_{a_2, x_2}^{\mathcal{A}_2} \mathcal{M}_{\text{unitary}}^{\mathcal{A}_2} \left(\mathcal{E}^1 \left(\pi_{a_1, x_1}^{\mathcal{A}_1} \mathcal{M}_{\text{unitary}}^{\mathcal{A}_1} (\rho_0) \pi_{a_1, x_1}^{\mathcal{A}_1} \right) \right) \pi_{a_2, x_2}^{\mathcal{A}_2} \right) \dots \pi_{a_{K-1}, x_{K-1}}^{\mathcal{A}_{K-1}} \right), \end{aligned} \quad (\text{E5})$$

where $K := \max(\text{OUT})$. We see that this prediction does not depend on any channel $\mathcal{E}^j$, setting x_j , nor any other property of an agent A_j for which $j > K$.

Using the definition of conditional probability, we can now derive an expression for setting-conditioned predictions (Definition 8), i.e., for the probability of a set of observers $\{A_{j_1}, A_{j_2}, \dots, A_{j_p}\}$ obtaining outcomes $\{a_{j_1}, a_{j_2}, \dots, a_{j_p}\}$ given $\{A_{l_1}, A_{l_2}, \dots, A_{l_q}\}$ measurement outcomes $\{a_{l_1}, a_{l_2}, \dots, a_{l_q}\}$ and setting $\vec{x}$. When using the definition of conditional probability, Eq. (E3) and simplifying by means of Eq. (E5), we obtain

$$P(a_{j_1}, a_{j_2}, \dots, a_{j_p} | a_{l_1}, \dots, a_{l_q}, \vec{x}) := \sum_{\substack{a_1 \text{ if } x_1=1 \& 1 \notin \text{OUT} \cup \text{IN} \\ a_2 \text{ if } x_2=1 \& 2 \notin \text{OUT} \cup \text{IN} \\ \vdots \\ a_{Q-1} \text{ if } x_{Q-1}=1 \& Q-1 \notin \text{OUT} \cup \text{IN}}} \frac{\text{numerator}}{\sum_{\substack{a_1 \text{ if } x_1=1 \\ a_2 \text{ if } x_2=1 \\ \vdots \\ a_{L-1} \text{ if } x_{L-1}=1}} \text{denominator}}, \quad (\text{E6})$$

numerator :=

$$\text{tr} \left[\pi_{a_Q, 1}^{\mathcal{A}_Q} \mathcal{M}_{\text{unitary}}^{\mathcal{A}_Q} \left(\mathcal{E}^{Q-1} \left(\pi_{a_{Q-1}, x_{Q-1}}^{\mathcal{A}_{Q-1}} \dots \mathcal{E}^2 \left(\pi_{a_2, x_2}^{\mathcal{A}_2} \mathcal{M}_{\text{unitary}}^{\mathcal{A}_2} \left(\mathcal{E}^1 \left(\pi_{a_1, x_1}^{\mathcal{A}_1} \mathcal{M}_{\text{unitary}}^{\mathcal{A}_1} (\rho_0) \pi_{a_1, x_1}^{\mathcal{A}_1} \right) \right) \pi_{a_2, x_2}^{\mathcal{A}_2} \right) \dots \pi_{a_{Q-1}, x_{Q-1}}^{\mathcal{A}_{Q-1}} \right) \right) \right], \quad (\text{E7})$$

$$\text{tr} \left[\pi_{a_Q, 1}^{\mathcal{A}_Q} \mathcal{M}_{\text{unitary}}^{\mathcal{A}_Q} \left(\mathcal{E}^{Q-1} \left(\pi_{a_{Q-1}, x_{Q-1}}^{\mathcal{A}_{Q-1}} \dots \mathcal{E}^2 \left(\pi_{a_2, x_2}^{\mathcal{A}_2} \mathcal{M}_{\text{unitary}}^{\mathcal{A}_2} \left(\mathcal{E}^1 \left(\pi_{a_1, x_1}^{\mathcal{A}_1} \mathcal{M}_{\text{unitary}}^{\mathcal{A}_1} (\rho_0) \pi_{a_1, x_1}^{\mathcal{A}_1} \right) \right) \pi_{a_2, x_2}^{\mathcal{A}_2} \right) \dots \pi_{a_{Q-1}, x_{Q-1}}^{\mathcal{A}_{Q-1}} \right) \right) \right], \quad (\text{E8})$$

denominator := (E9)

$$\text{tr} \left[\pi_{a_L, l}^{A_L} \mathcal{M}_{\text{unitary}}^{A_L} \left(\mathcal{E}^{L-1} \left(\pi_{a_{L-1}, x_{L-1}}^{A_{L-1}} \dots \mathcal{E}^2 \left(\pi_{a_2, x_2}^{A_2} \mathcal{M}_{\text{unitary}}^{A_2} \left(\mathcal{E}^1 \left(\pi_{a_1, x_1}^{A_1} \mathcal{M}_{\text{unitary}}^{A_1} (\rho_0) \pi_{a_1, x_1}^{A_1} \right) \right) \pi_{a_2, x_2}^{A_2} \right) \dots \pi_{a_{L-1}, x_{L-1}}^{A_{L-1}} \right) \right) \right], \quad (\text{E10})$$

where $Q := \max(\text{OUT} \cup \text{IN})$, $\text{IN} = \{l_1, l_2, \dots, l_q\}$, and $L := \max(\text{IN})$. Furthermore, we have the constraint that $x_{j_1} = x_{j_2} = \dots = x_{j_p} = x_{l_1} = x_{l_2} = \dots = x_{l_q} = 1$, since we are reasoning about these measurement outcomes. Note that if we now sum $P(a_{j_1}, a_{j_2}, \dots, a_{j_p} | a_{l_1}, \dots, a_{l_q}, \vec{x})$ over the elements in the set $\{a_{j_1}, a_{j_2}, \dots, a_{j_p}\}$, we obtain one, and, thus, the distribution is normalized. Notice also that we are not summing over outcomes a_j for which $x_j = 0$, since these correspond to the case said measurement is modeled unitarily. Also, as mentioned previously, we have also preselected state ρ_0 . The “postselected state” is merely the postmeasurement state of the last measurement performed by observers A_j , $j \in \text{OUT}$. We can easily derive an expression for our predictions (Definition 2) using Eq. (E6) and our prior $P(\vec{x})$:

$$P(a_{j_1}, a_{j_2}, \dots, a_{j_p} | a_{l_1}, \dots, a_{l_q}) \\ := \sum_{\vec{x}} P(a_{j_1}, a_{j_2}, \dots, a_{j_p} | a_{l_1}, \dots, a_{l_q}, \vec{x}) P(\vec{x}), \quad (\text{E11})$$

where for consistency we have defined $P(a_{j_1}, a_{j_2}, \dots, a_{j_p} | a_{l_1}, \dots, a_{l_q}, \vec{x}) = 0$ if there exists $k \in \{j_1, \dots, j_p, l_1, \dots, l_q\}$ such that $a_k \neq \perp$ and $x_k = 0$.

The simplification coming from the iterative application of Eq. (E3) to our conditional probability has important physical consequences, namely, that the expressions (E6) and (E11) readily do not depend on the channels, measurement schemes, nor settings of agents in the future of when the agents in $\{A_k | k \in \text{OUT} \cup \text{IN}\}$ perform their

operations, i.e., in the future of the agents whose measurement outcomes are being reasoned about.

APPENDIX F: REDUCTION OF THE AUGMENTED CIRCUIT IN STANDARD QUANTUM SCENARIOS

In Sec. VI, we defined the subclass of EWFSs which we call standard quantum scenarios (Definition 13) and showed that objective, setting-independent predictions emerge in this case. Here, we show that, in EWFSs corresponding to standard quantum scenarios, the augmented circuit of our framework reduces to a standard form quantum circuit without the “Heisenberg-cut” settings. We define two forms of standard quantum circuit representations below before proving this.

Definition F1 (standard quantum circuit representations). Consider a quantum protocol involving N agents $A := \{A_1, \dots, A_N\}$ and m systems $S := \{S_1, \dots, S_m\}$ where each agent A_i performs a projective measurement $\mathcal{M}^{A_i} := \{\pi_{a_i}^{S_i}\}_{a_i \in O_i}$ at time t_i that acts nontrivially on a subset $S_i \subseteq S$ of the systems, obtaining an outcome a_i that can take values in a set O_i , followed by a channel $\mathcal{E}_i$ that may act on all S . A standard quantum circuit representation of such a protocol corresponds to a circuit of one of the two following types: In one case, all measurements are modeled as projectors on the systems S alone, and, in another case, all measurements can be equivalently purified to unitaries on the systems and some ancillas.

- (1) *C^{sys}-form quantum circuit (Fig. 7)*—A quantum circuit acting on S which is defined through the

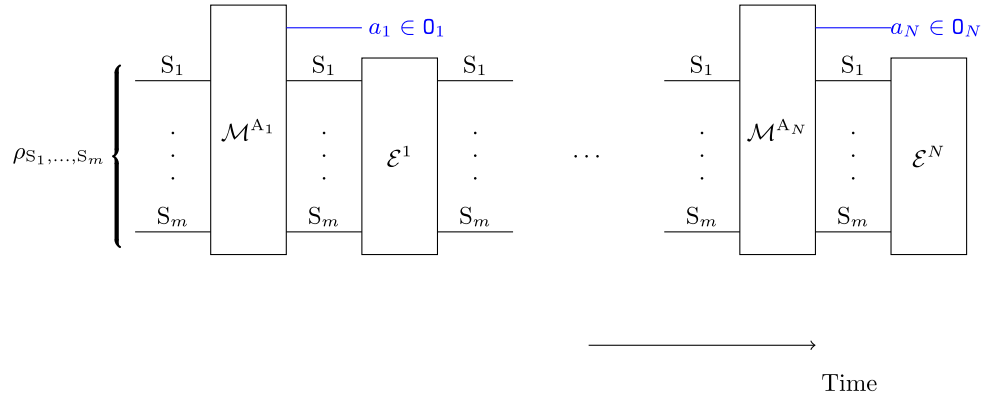


FIG. 7. A C^{sys} -form standard quantum circuit. Here, O_i is the set of possible nontrivial values of the outcome a_i . Each $\mathcal{M}^{A_i} := \{\pi_{a_i}^{S_i}\}_{a_i}$ implements a projective measurement of the subset $S_i \subseteq S := \{S_1, \dots, S_m\}$ of systems.

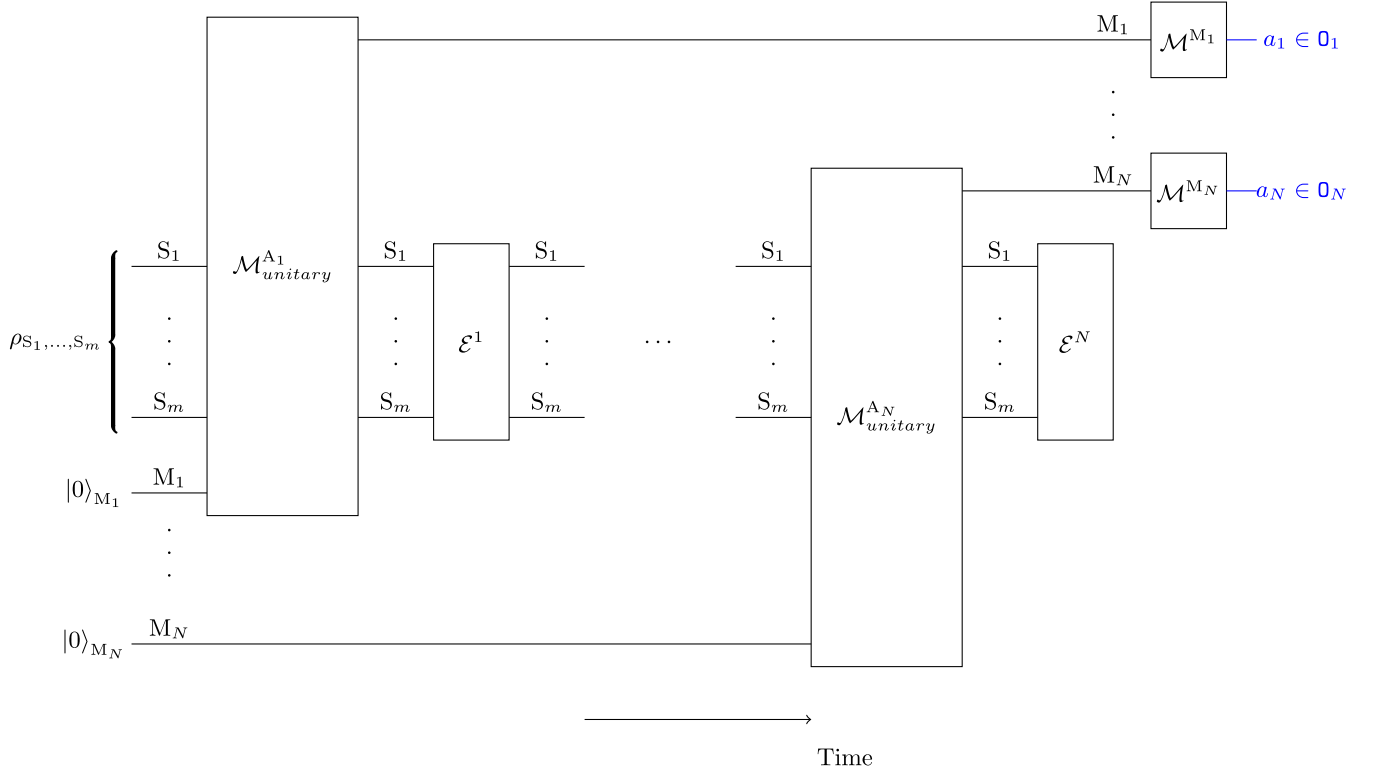


FIG. 8. A $C^{\text{sys+anc}}$ -form standard quantum circuit where each measurement in Fig. 7 is purified to a unitary using an ancilla. Here, $\mathcal{O}_i$ is the set of possible nontrivial values of the outcome a_i . The measurements $\mathcal{M}^{M_i} := \{\pi_{a_i}^{M_i}\}_{a_i}$ applied in the global future implement the isomorphic projective measurement of the ancilla M_i , as do the measurements $\mathcal{M}^{A_i}$ (from Fig. 7) on the subset $S_i \subseteq S := \{S_1, \dots, S_m\}$ of systems. It is immediate to see (and well known) that the present circuit and that in Fig. 7 are operationally equivalent.

composition $\mathcal{E}_N \circ \mathcal{M}_{unitary}^{A_N} \circ \dots \circ \mathcal{E}_1 \circ \mathcal{M}_{unitary}^{A_1}$, where each operation is defined over all of S but it is given that each $\mathcal{M}^{A_i}$ acts nontrivially on some subset $S_i \subseteq S$. Outcome probabilities are calculated by applying the Born rule to the circuit, with the projective measurements $\mathcal{M}^{A_i} := \{\pi_{a_i}^{S_i}\}_{a_i \in \mathcal{O}_i}$ on S_i .

- (2) $C^{\text{sys+anc}}$ -form quantum circuit (Fig. 8)—A quantum circuit acting on $S \cup \{M_1, \dots, M_N\}$, where M_i denotes an ancillary quantum system corresponding to the measurement $\mathcal{M}^{A_i}$ whose state space is isomorphic to that of the systems $S_i \subseteq S$ on which the measurement acts nontrivially. It is defined through the composition $\mathcal{E}_N \circ \mathcal{M}_{unitary}^{A_N} \circ \dots \circ \mathcal{E}_1 \circ \mathcal{M}_{unitary}^{A_1}$, where each measurement $\mathcal{M}^{A_i}$ is purified to a unitary interaction $\mathcal{M}_{unitary}^{A_i}$ acting on $S \cup M_i$ (and nontrivially on $S_i \cup M_i$), where $\mathcal{M}_{unitary}^{A_i}$ corresponds to the unitary that implements a coherent copy from S_i to M_i in the orthonormal basis given by the measurement projectors [as defined in Eq. (3)]. Each $\mathcal{E}_i$ acts only on S . Outcome probabilities are calculated by measuring the ancillas M_i (using isomorphic projectors $\{\pi_{a_i}^{M_i} := |a_i\rangle\langle a_i|_{M_i}\}_{a_i \in \mathcal{O}_i}$) at a time

$t_f > t_N$ at the end of the protocol (using the Born rule).

These two forms of circuits are illustrated in Figs. 7 and 8.

Theorem F1 (recovering standard quantum circuits). If an EWFS corresponds to a standard quantum scenario (Definition 13), then its augmented circuit can be equivalently reduced to a standard quantum circuit, such that the same (nontrivial) predictions are obtained from the original augmented circuit, the C^{sys} -form standard circuit, or the $C^{\text{sys+anc}}$ -form standard circuit. Explicitly, for any disjoint sets $\vec{a}_j = (a_{j_1}, \dots, a_{j_p})$ and $\vec{a}_l = (a_{l_1}, \dots, a_{l_q})$ of outcomes, and any choice of settings $\vec{x} = \vec{\xi}$ such that $x_i = 1$ for all $i \in \{j_1, \dots, j_p, l_1, \dots, l_q\}$, we have

$$P_{\text{aug}}(\vec{a}_j = \vec{a}_j | \vec{a}_j = \vec{a}_j, \vec{x} = \vec{\xi}) = P_{\text{std}}(\vec{a}'_j = \vec{a}_j | \vec{a}'_j = \vec{a}_j), \quad (\text{F1})$$

where the P_{aug} refers to setting-conditioned predictions in the augmented circuit of the EWFS and P_{std} refers to predictions in an equivalent C^{sys} -form or $C^{\text{sys+anc}}$ -form standard quantum circuit (where no settings are involved).

A proof of this theorem can be found in Appendix J.

APPENDIX G: DETAILED ANALYSIS OF THE FR SCENARIO

1. Entanglement version of the FR scenario

In the main text, we provided a brief overview of the entanglement version of the FR scenario as well as a simple explanation of our resolution of the paradox. Having reviewed this scenario in detail in Appendix D 2, we now provide a more detailed analysis of the same, showing the explicit calculation for every prediction involved.

We reproduce each of the individual statements used in the reasoning of the FR scenario described in Appendix D 2, i.e., those captured by Eqs. (D4) and (D6). Additionally, this reasoning occurs only in a round where the superagents observe the outcomes $u = w = \text{ok}$, which is associated with a probability $\frac{1}{12}$ in FR's arguments as shown in Eq. (D3). We start by reproducing this probability as an explicit setting conditioning prediction in our framework. The probability of Eq. (D3) is equivalent to conventional prediction $P_{\text{conv}}^{\text{FR}}(u = w = \text{ok})$ (Sec. III B) of the FR scenario; in computing this, FR apply the U assumption implicitly to model the measurements of the agents Alice and Bob purely unitarily [using $\mathcal{M}_{\text{unitary}}^A$ and $\mathcal{M}_{\text{unitary}}^B$ as seen in Eq. (D3)]. This corresponds precisely to the setting choices $x_A = x_B = 0$ (as expected from the general mapping from conventional to setting-conditioned predictions given in Appendix J 2). Therefore, it follows immediately that

$$\begin{aligned} P(u = w = \text{ok} | |\psi\rangle^{t=2}) &:= P_{\text{conv}}^{\text{FR}}(u = w = \text{ok}) \\ &= P(u = w = \text{ok} | x_A = x_B = 0). \end{aligned} \quad (\text{G1})$$

Note that, in our notation for predictions, we do not explicitly condition on the initial state of the scenario, as this is taken to be in the common knowledge of all agents. See Sec. VII A for a discussion on how our framework and arguments can generalize to the case where one relaxes this common knowledge assumption.

Now we proceed to analyzing each of the statements that agents make when the above postselection on $u = w = \text{ok}$ succeeds. Consider the statement obtained in Eq. (D4); here, Ursula, upon knowing that $u = w = \text{ok}$, reasons about Bob's outcome b using the state $|\psi^{t=2}\rangle_{\text{RASB}}$ and concludes that $u = w = \text{ok} \Rightarrow b = 1$. This logical statement is equivalently expressed in probabilities through the conventional prediction $P_{\text{conv}}(b = 1 | u = w = \text{ok}) = 1$.

We can readily extract the setting choices implicit in the calculation of this probability for the FR protocol. Note that $|\psi^{t=2}\rangle$ is obtained by applying $M_{\text{unitary}}^A \otimes M_{\text{unitary}}^B$ to the initial state $|\psi^{t=1}\rangle$, and in calculating the above-mentioned probability for $b = 1$ using the Born rule, FR apply the projector $\pi_{1,1}^B = |11\rangle\langle 11|_{\text{SB}}$ to $|\psi^{t=2}\rangle$. This precisely corresponds to the assigning $x_2 = 1$ for Bob's setting. On the other hand, they model Alice's measurement as a purely unitary evolution (M_{unitary}^A) as seen by Ursula in this reasoning step; therefore, the setting choice used for Alice in this reasoning is $x_1 = 0$. Making these setting choices explicit, we see that this probability calculated in the FR reasoning is equivalent to the setting-conditioned prediction $P(b = 1 | u = w = \text{ok}, (x_1, x_2) = (0, 1)) = 1$ in our framework.

Indeed, one can calculate this prediction from the augmented circuit in Fig. 5 for the FR protocol (using the Born rule and the well-known rule for conditional probabilities) and would obtain the same. We demonstrate this below; for further details on how the probability rule for setting-conditioned predictions in augmented circuits is derived, see Appendix E:

$$\begin{aligned} P(b = 1 | u = w = \text{ok}, (x_1, x_2) = (0, 1)) \\ = \frac{P(b = 1, u = w = \text{ok} | (x_1, x_2) = (0, 1))}{P(u = w = \text{ok} | (x_1, x_2) = (0, 1))}. \end{aligned} \quad (\text{G2})$$

That this expression evaluates to unit probability is evident from the following calculation of the numerator and denominator of this expression for the FR protocol:

$$\begin{aligned} P(b = 1, u = w = \text{ok} | (x_1, x_2) = (0, 1)) &= \frac{1}{12} \\ &= \left| \langle \text{ok} |_{\text{RA}} \otimes \langle \text{ok} |_{\text{SB}} \left(1_{\text{RA}} \otimes \pi_{x_2=1, b=1}^B \right) | \psi^{t=2} \rangle \right|^2 \\ &= \left| \langle \text{ok} |_{\text{RA}} \otimes \langle \text{ok} |_{\text{SB}} \left(1_{\text{RA}} \otimes |11\rangle\langle 11|_{\text{SB}} \right) | \psi^{t=2} \rangle \right|^2, \end{aligned} \quad (\text{G3})$$

$$\begin{aligned} P(u = w = \text{ok} | (x_1, x_2) = (0, 1)) &= \frac{1}{12} \\ &= \sum_{b \in \{0,1\}} \left| \langle \text{ok} |_{\text{RA}} \otimes \langle \text{ok} |_{\text{SB}} \left(1_{\text{RA}} \otimes \pi_{x_2=1, b}^B \right) | \psi^{t=2} \rangle \right|^2 \\ &= \sum_{b \in \{0,1\}} \left| \langle \text{ok} |_{\text{RA}} \otimes \langle \text{ok} |_{\text{SB}} \left(1_{\text{RA}} \otimes |bb\rangle\langle bb|_{\text{SB}} \right) | \psi^{t=2} \rangle \right|^2. \end{aligned} \quad (\text{G4})$$

Having formalized FR's logical statement $u = w = \text{ok} \Rightarrow b = 1$ as the setting-conditioned prediction $P(b = 1|u = w = \text{ok}, (x_1, x_2) = (0, 1)) = 1$ in our framework, we obtain the corresponding explicit version of the statement:

$$u = w = \text{ok} \wedge (x_1, x_2) = (0, 1) \Rightarrow b = 1. \quad (\text{G5})$$

We now proceed to the next statement of the FR reasoning, given by Eq. (D5), where Ursula reasons about Bob's reasoning of Alice through the statement $b = 1 \Rightarrow a = 1$. This is equivalently expressed in terms of probabilities through the conventional prediction $P_{\text{conv}}(a = 1|b = 1) = 1$.

To evaluate this probability using the Born rule as FR do, we must apply the projector $\pi_{1,1}^A \otimes \pi_{1,1}^B = |11\rangle\langle 11|_{\text{RA}} \otimes$

$|11\rangle\langle 11|_{\text{SB}}$ to $|\psi^{t=2}\rangle_{\text{RASB}}$, and it is evident that the implicit setting choices for Alice and Bob used here are $(x_1, x_2) = (1, 1)$. Therefore, making this explicit, we have $P(a = 1|b = 1, (x_1, x_2) = (1, 1)) = 1$. We can again verify this from the probability rule for our augmented circuit (which is simply the Born rule and standard conditional probability rule):

$$\begin{aligned} P(a = 1|b = 1, (x_1, x_2) = (1, 1)) \\ = \frac{P(a = 1, b = 1|(x_1, x_2) = (1, 1))}{P(b = 1|(x_1, x_2) = (1, 1))}. \end{aligned} \quad (\text{G6})$$

That this evaluates to unit probability is immediate from the following expressions for the numerator and denominator:

$$\begin{aligned} P(a = 1, b = 1|(x_1, x_2) = (1, 1)) &= \frac{1}{3} \\ &= \left| \left(\pi_{x_1=1, a=1}^A \otimes \pi_{x_2=1, b=1}^B \right) |\psi^{t=2}\rangle \right|^2 \\ &= \left| \left(|11\rangle\langle 11|_{\text{RA}} \otimes |11\rangle\langle 11|_{\text{SB}} \right) |\psi^{t=2}\rangle \right|^2, \end{aligned} \quad (\text{G7})$$

$$\begin{aligned} P(b = 1|(x_1, x_2) = (1, 1)) &= \frac{1}{3} \\ &= \left| \left(1_{\text{RA}} \otimes \pi_{x_2=1, b=1}^B \right) |\psi^{t=2}\rangle \right|^2 \\ &= \left| \left(1_{\text{RA}} \otimes |11\rangle\langle 11|_{\text{SB}} \right) |\psi^{t=2}\rangle \right|^2. \end{aligned} \quad (\text{G8})$$

From this, as before, we can extract the explicit version of the logical statement:

$$b = 1 \wedge (x_1, x_2) = (1, 1) \Rightarrow a = 1. \quad (\text{G9})$$

We now turn to the third statement of the FR reasoning given in Eq. (D6), where Ursula reasons about Bob's reasoning about Alice's reasoning about Wigner, through the statement $a = 1 \Rightarrow w = \text{fail}$. This equivalent probabilistic version is given by the conventional prediction $P_{\text{conv}}(w = \text{fail}|a = 1) = 1$.

Analyzing how FR calculate this probability using the Born rule, we see that this involves applying the projector $\pi_{1,1}^A = |11\rangle\langle 11|_{\text{RA}}$ to the state $|\psi^{t=2}\rangle$ which gives us Alice's setting $x_1 = 1$. Moreover, Bob is modeled purely unitarily here, through M_{unitary}^B , and we have $x_2 = 0$. Therefore, the explicit setting-conditioned prediction corresponding to this reasoning step of FR is $P(w = \text{fail}|a = 1, (x_1, x_2) = (1, 0)) = 1$. This can be verified within our framework as follows:

$$P(w = \text{fail}|a = 1, (x_1, x_2) = (1, 0)) = \frac{P(w = \text{fail}, a = 1|(x_1, x_2) = (1, 0))}{P(a = 1|(x_1, x_2) = (1, 0))}. \quad (\text{G10})$$

The numerator and denominator are evaluated below, which makes it evident that the expression above evaluates to unity:

$$\begin{aligned} P(w = \text{fail}, a = 1|(x_1, x_2) = (1, 0)) &= \frac{2}{3} \\ &= \left| 1_{\text{RA}} \otimes \langle \text{fail} |_{\text{SB}} \left(\pi_{x_1=1, a=1}^A \otimes 1_{\text{SB}} \right) |\psi^{t=2}\rangle \right|^2 \\ &= \left| 1_{\text{RA}} \otimes \langle \text{fail} |_{\text{SB}} \left(|11\rangle\langle 11|_{\text{RA}} \otimes 1_{\text{SB}} \right) |\psi^{t=2}\rangle \right|^2, \end{aligned} \quad (\text{G11})$$

$$\begin{aligned}
P(a=1|(x_1, x_2) = (1, 0)) &= \frac{2}{3} \\
&= |(\pi_{x_1=1, a=1}^A \otimes 1_{SB})|\psi^{t=2}\rangle|^2 \\
&= |(|11\rangle\langle 11|_{RA} \otimes 1_{SB})|\psi^{t=2}\rangle|^2.
\end{aligned} \tag{G12}$$

Then, the corresponding, explicit version of the logical statement is

$$a = 1 \wedge (x_1, x_2) = (1, 0) \Rightarrow w = \text{fail}. \tag{G13}$$

Therefore, we have explicitly derived all the statements in Table I while highlighting the setting choices implicit in each of FR's statements. As we have seen, making explicit these setting choices is sufficient to resolve the apparent paradox.

2. Prepare-and-measure version of the FR scenario

We have reviewed the original prepare and measure of the FR scenario in Appendix D 3. Here, we show that the resolution proposed for the entanglement version is also applicable to the original (prepare-and-measure) version of the FR paradox, Ref. [3]. We first show explicitly how original version of the FR protocol can also be modeled as an augmented circuit. Then, we go through the reasoning of the prepare-and-measure version on a statement-by-statement basis of their apparent proof of their paradox. For every statement, we reveal the different settings which said

statements are contingent on but not stated by the authors of the FR paper.

a. The augmented circuit

The augmented circuit of the original prepare-and-measure version of the FR protocol is given in Fig. 9. An equivalent version of this circuit is given in Fig. 10, that makes the mapping to the entanglement version of the FR scenario more explicit.

The circuits encode the states and measurements of the original protocol in the computational basis as follows. The initial state $\sqrt{\frac{1}{3}}|\text{heads}\rangle_R + \sqrt{\frac{2}{3}}|\text{tails}\rangle_R$ of the coin in $\bar{F}$'s lab [Eq. (D8)] is represented in the computational basis as $|\psi\rangle_R = \sqrt{\frac{1}{3}}|0\rangle_R + \sqrt{\frac{2}{3}}|1\rangle_R$; all other systems ($\bar{L}$, S , and L) are initialized to $|0\rangle$. Then, the measurement of $\bar{F}$ corresponds to a computational basis measurement, with the outcome $r = \text{heads}$ identified with $r = 0$ and $r = \text{tails}$ identified with $r = 1$. The preparation of S carried out by $\bar{F}$ based on the outcome of their measurement on R corresponds to a controlled Hadamard with the states $|\downarrow\rangle_S$ and $|\uparrow\rangle_S$ of the FR scenarios represented as $|0\rangle_S$ and $|1\rangle_S$ here. Similarly, the measurement of F then also becomes a computational basis measurement with the outcome $z = +\frac{1}{2}$ identified with $z = 0$ and $z = -\frac{1}{2}$ identified with $z = 1$. The projectors $\Pi_{x_1}^{\bar{F}}$ and $\Pi_{x_2}^F$ acting on the systems $R\bar{L}$ and SL , respectively, are identical to the projectors $\Pi_{x_1}^A$ and $\Pi_{x_2}^B$ in Eq. (10) acting on the systems RA and SB , and the final measurements of $\bar{W}$ and W are the

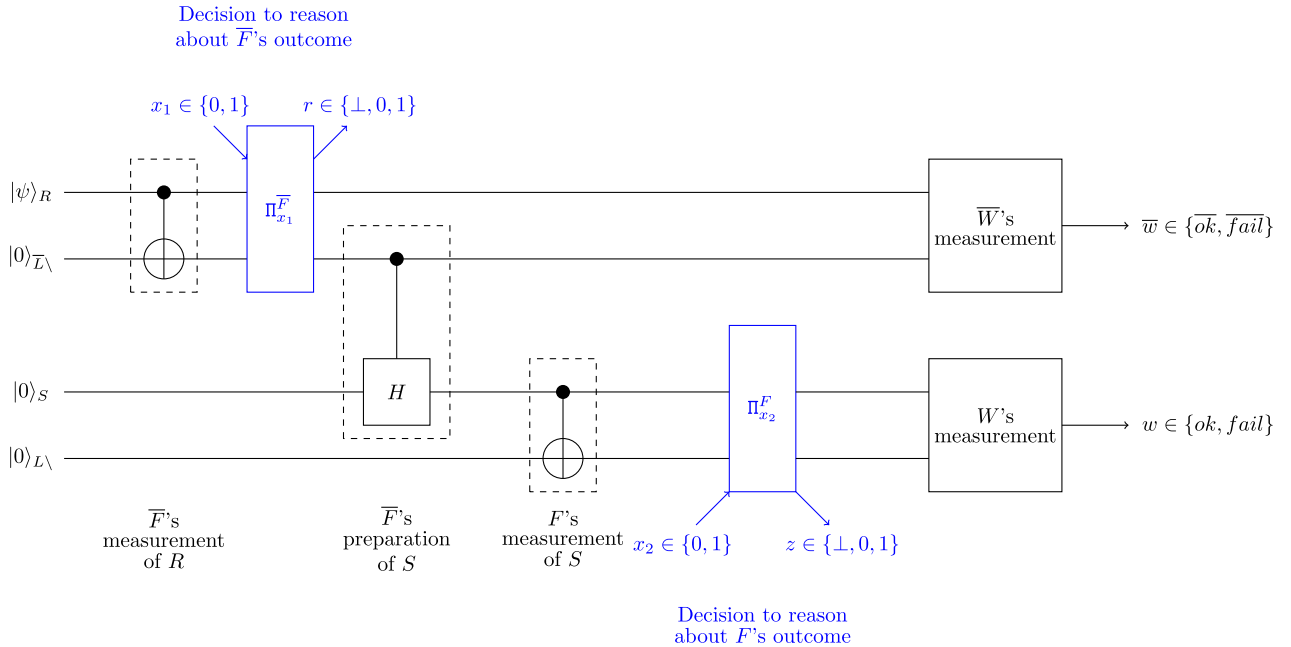


FIG. 9. Augmented circuit for the prepare-and-measure version of the FR scenario expressed in terms of the computational basis, as explained in the main text.

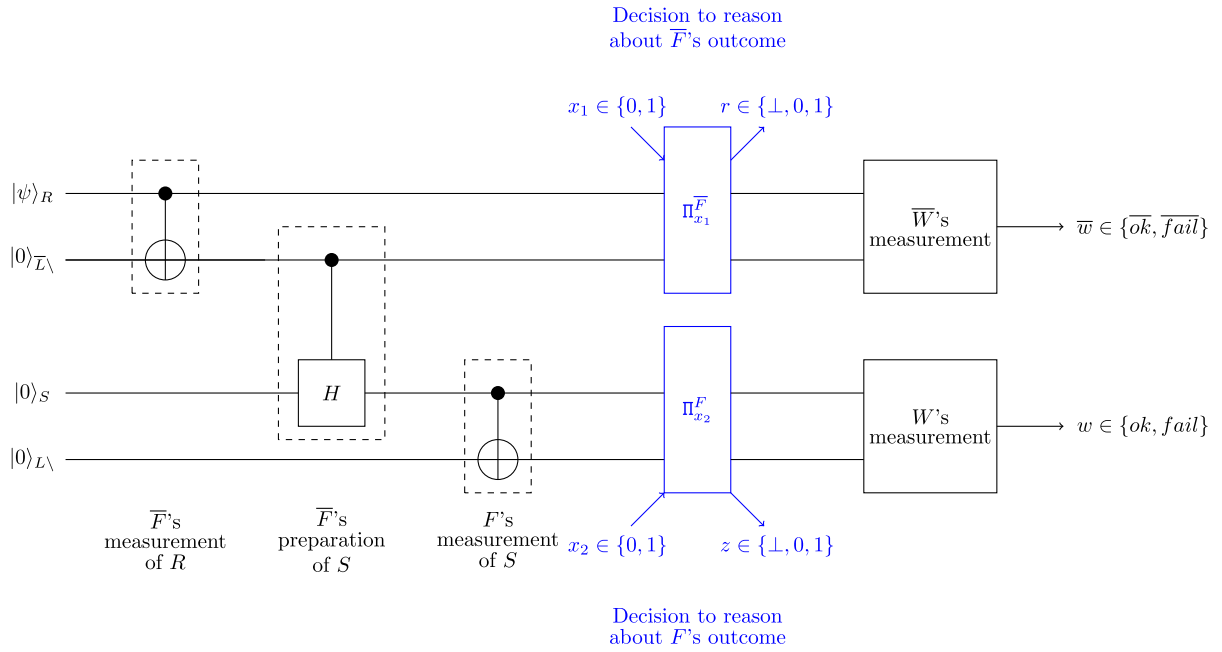


FIG. 10. Equivalent version of the augmented circuit (Fig. 9) of the prepare-and-measure FR protocol. Note that the blue box corresponding to $\Pi_{x_1}^{\bar{F}}$ commutes with the controlled Hadamard gate, since the control is on the same basis as the measurements associated with the box. Then, it is easy to verify that the joint state of $R\bar{L}\backslash SL$ just before the blue boxes (i.e., just after F 's measurement of S) is precisely the same state as $|\psi^{t=2}\rangle_{RASB}$ [Eq. (D2)] of the entanglement version of the FR scenario, with $\bar{L}\backslash$ and $L\backslash$ playing the role of A and B , respectively.

same as the entanglement version of Appendix D 2, i.e., $\{|\text{ok}\rangle_{SL\backslash} = (1/\sqrt{2})(|00\rangle - |11\rangle)_{SL\backslash}, |\text{fail}\rangle_{SL\backslash} := (1/\sqrt{2}) \times (|00\rangle + |11\rangle)_{SL\backslash}\}$ and similarly for $\{|\overline{\text{ok}}\rangle_{R\bar{L}\backslash}, |\overline{\text{fail}}\rangle_{R\bar{L}\backslash}\}$.

b. Statement-by-statement analysis

We now analyze their constructive proof step by step and point out what settings are required to reproduce their statements. We use a notation which is close to, but not identical to, that of the authors to aid comparison.

The authors use the following notation to denote the statements by specific agents at particular times: $G^t\text{：“}k\text{”}$ where $G \in \{W, \bar{W}, F, \bar{F}\}$ denotes the agent G making statement k at time t . Here, x_1 are for the measurements of $\bar{F}$ while x_2 are for F 's. Analogously to in the entanglement scenario, we do not need to consider settings for W nor $\bar{W}$, since they are superagents. We use italics when referring to the reasoning of the authors. What is more, we use the same notation as the authors to specify different statements, but with an additional subscript indicating the (x_2, x_1) setting that said observer is using when making the statement in the corresponding augmented circuit: $G^t_{(x_1, x_2)}\text{：“}k\text{”}$. The setting $(0, 0)$ corresponds to the case where the agent is not reasoning about F nor $\bar{F}$'s measurement outcome. Note that the reasoning of a particular observer depends on only their choice of (x_1, x_2) settings and not on those of the other observers. Furthermore, we

say $G^t_{(x_1, x_2)}$ is correct if G^t holds under settings (x_1, x_2) , $G^t_{(x_1, x_2)}$ is false if statement G^t does not hold under settings (x_1, x_2) , and $G^t_{(x_1, x_2)} = \emptyset$ if statement G^t involves reasoning about an observer's measurement outcome when the choice (x_1, x_2) does not allow observer G to reason about said observer.

The first statement is made by $\bar{F}$ at time $n:00$: $\bar{F}^{n:00}$: “The value w is obtained by a measurement of L with respect to basis $\{\pi_{w=\text{ok}}^{n:10}, \pi_{w=\text{fail}}^{n:10}\}$, which is completed at time $n:31$ ”. Here, $\pi_{w=\text{ok}}^{n:10}$ and $\pi_{w=\text{fail}}^{n:10}$ are the projectors onto the $|\text{ok}\rangle_{L\backslash}, |\text{fail}\rangle_{L\backslash}$ basis. This statement is correct according to the framework of this paper for all (x_1, x_2) , i.e., $\bar{F}^{n:00}_{(0,0)} = \bar{F}^{n:00}_{(0,1)} = \bar{F}^{n:00}_{(1,0)} = \bar{F}^{n:00}_{(1,1)} = \text{correct}$.

Then, if $\bar{F}$ got $r = \text{tails}$ in her measurement of the coin flip, she would make the statement $\bar{F}^{n:01}$: “The spin S is in state $|\rightarrow\rangle_S$ at time $n:10$ ”. In our formalism, this statement would also hold, since $\bar{F}$ is reasoning about only the measurement she made; we, thus, have $\bar{F}^{n:01}_{(1,0)} = \bar{F}^{n:01}_{(1,1)} = \text{correct}$ and $\bar{F}^{n:01}_{(0,0)} = \bar{F}^{n:01}_{(0,1)} = \emptyset$. Any further statements made by agent $\bar{F}$ would have to be preselected on the spin being in state $|\rightarrow\rangle_S$. In Fig. 6, this would correspond to the selection of branch 1 and multiplying it by $\sqrt{3/2}$ to renormalize the branch. The authors then go on to make the following claim: $\bar{F}^{n:00}$ and $\bar{F}^{n:01}$ inserted into Q imply $w = \text{fail}$ (this is claim $\bar{F}^{n:02}$ in Table 3). This statement holds only if one does not take into account F 's measurement at

time $n:10$. In other words, $\bar{F}^{n:02}_{(1,0)} = \text{correct}$, $\bar{F}^{n:02}_{(1,1)} = \text{false}$. To see this, note from Fig. 6 that their conclusion follows from noting the cancellation of the two blue $|\text{ok}\rangle_L$ kets when we preselect on branch 2. However, branch 2 is further split into subbranches 2.1.1 and 2.1.2 via F 's measurement. This splitting causes the blue colored kets to not cancel each other out from $\bar{F}$'s perspective when reasoning under settings $(x_1, x_2) = (1, 1)$. In terms of equations, under settings $(x_1, x_2) = (1, 1)$ we would conclude that $\bar{F}^{n:00}$ and $\bar{F}^{n:01}$ imply that the probability that $w = \text{fail}$ is

$$\begin{aligned} P(w = \text{fail} | r = \text{tails}, (x_1, x_2) = (1, 1)) \\ = \sum_{z \in \{-1/2, 1/2\}} P(w = \text{fail}, z | r = \text{tails}, (x_1, x_2) = (1, 1)) \\ = \frac{\sum_{z \in \{-1/2, 1/2\}} \text{tr} \left[\pi_{\text{fail}}^W \pi_z^F \pi_{\text{tail}}^{\bar{F}} \rho_{\text{Fi}} \pi_{\text{tail}}^{\bar{F}} \pi_z^F \right]}{\sum_{z \in \{-1/2, 1/2\}} \text{tr} \left[\pi_w^W \pi_z^F \pi_{\text{tail}}^{\bar{F}} \rho_{\text{Fi}} \pi_{\text{tail}}^{\bar{F}} \pi_z^F \right]} \\ = \frac{\sum_{z \in \{-1/2, 1/2\}} \text{tr} \left[\pi_{\text{fail}}^W \pi_z^F \pi_{\text{tail}}^{\bar{F}} \rho_{\text{Fi}} \pi_{\text{tail}}^{\bar{F}} \pi_z^F \right]}{\sum_{z \in \{-1/2, 1/2\}} \text{tr} \left[\pi_{\text{tail}}^{\bar{F}} \rho_{\text{Fi}} \right]} = \frac{1}{2}, \end{aligned} \quad (\text{G14})$$

where $\rho_{\text{Fi}} := U_F U_{\text{spin}} U_{\bar{F}} \rho_0 (U_F U_{\text{spin}} U_{\bar{F}})^\dagger$ with $\rho_0 := |\psi_0\rangle \times \langle\psi_0|$, $|\psi_0\rangle$ being the initial state [i.e., lhs of Eq. (D11)]. The unitaries U_F , U_{spin} , and $U_{\bar{F}}$ are those of the protocol (see Fig. 6) and give rise to the final state $\rho_{\text{Fi}} = |\psi_{\text{Fi}}\rangle \langle\psi_{\text{Fi}}|$ with $|\psi_{\text{Fi}}\rangle$ given by the rhs of Eq. (D11). F 's measurement outcome at time $n:00$ is denoted by r (with a denoting the corresponding random variable) and $\{\pi_z^F\}_{z \in \{-1/2, 1/2\}}$ is the projective measurement of F 's measurement at time $n:10$. We note that, in the notation of the general framework of Sec. III, these projectors would be explicitly written as $\{\pi_{1,z}^F\}_{z \in \{-1/2, 1/2\}}$, as they correspond to the case of choosing the setting to be 1 for that measurement. In order to avoid clutter, here and in the following, we drop the setting subscript "1" in all such projectors, as the meaning is evident from the context of the protocol at hand. In Eq. (G14), we have used the fact that the unitary transformations taking us from the initial state to the final state commute with the measurement projectors; i.e., we have used the equivalence between the circuits in Figs. 9 and 10.

Since this probability in Eq. (G14) is less than one, we can conclude that $\bar{F}^{n:02}_{(1,1)} = \text{false}$.

Now it is F 's turn to make a statement. $F^{n:10}$: "The value z is obtained by a measurement of spin S with respect to $\{\pi_{z=-1/2}^{n:10}, \pi_{z=1/2}^{n:10}\}$, which is completed at time $n:11$ ". This statement clearly holds for all setting choices for F in which F can reason about her outcome, since it is merely stating one of the rules of the protocol: $F^{n:10}_{(0,1)} = F^{n:10}_{(1,1)} = \text{correct}$. They now further go on to state: Suppose now that F

observed $z = 1/2$ in round n . Since $\langle \downarrow | \pi_{z=-1/2}^{n:10} | \downarrow \rangle = 1$, it follows from Q that S was not in state $|\downarrow\rangle_S$ and, hence, that the random value r was not heads. Therefore, $F^{n:12}$: "I am certain that $\bar{F}$ knows that $r = \text{tails}$ at time $n:11$ ". Here, F is reasoning about both her measurement outcome and that of $\bar{F}$. Therefore, by definition, this statement makes sense only when F chooses setting $(x_1, x_2) = (1, 1)$, since, in the case $(x_1, x_2) = (0, 1)$, F cannot reason about $\bar{F}$'s measurement outcome since there is no classical outcome to assign to it. We, therefore, have $F^{n:12}_{(0,1)} = F^{n:12}_{(1,0)} = \emptyset$. Meanwhile, the following equation verifies that $F^{n:12}_{(1,1)} = \text{correct}$. Using Eq. (E6), if we postselect on F getting $z = 1/2$ in round n , then we find that the probability that $\bar{F}$ got tails is one:

$$\begin{aligned} P(r = \text{tails} | z = 1/2, (x_1, x_2) = (1, 1)) \\ = \frac{\text{tr} [\pi_{1/2}^F \pi_{\text{tail}}^{\bar{F}} \rho_{\text{Fi}} \pi_{\text{tail}}^{\bar{F}}]}{\sum_{r \in \{\text{tails}, \text{heads}\}} \text{tr} [\pi_{1/2}^F \pi_r^{\bar{F}} \rho_{\text{Fi}} \pi_r^{\bar{F}}]} = 1, \end{aligned} \quad (\text{G15})$$

where $\pi_{\text{tails}}^{\bar{F}} = |\bar{t}\rangle \langle \bar{t}|_L$ and $\pi_{\text{heads}}^{\bar{F}} = |\bar{h}\rangle \langle \bar{h}|_L$ are the projectors onto the lab of $\bar{F}$, corresponding to the two outcomes of the coin toss. The last equality follows from noting that $\text{tr} [\pi_{1/2}^F \pi_{\text{heads}}^{\bar{F}} \rho_{\text{Fi}} \pi_{\text{heads}}^{\bar{F}}] = 0$.

Similarly, the above equation can also be concluded from Fig. 6 by noting that if we are in branch 2.1.2 (this corresponds to F getting outcome $z = 1/2$), then the only measurement outcome of $\bar{F}$ which leads to this branch is $r = \text{tails}$.

The authors of the FR paradox then claim: Therefore, from $F^{n:12}$ and invoking Q , we conclude $F^{n:13}$: "I am certain that $\bar{F}$ is certain that W will observe $w = \text{fail}$ at time $n:31$." Statements $\{F^{n:13}_{(x_1, x_2)}\}_{x_1, x_2}$ do not correspond to a single statement in our framework, since $F^{n:13}$ is a concatenation of two statements "upon observing $z = 1/2$, I know with certainty that $r = \text{tails}$ " made by F and "upon observing $r = \text{tails}$, I know with certainty that $w = \text{fail}$ " made by $\bar{F}$. Note that these two statements are precisely $F^{n:12}$ and $\bar{F}^{n:02}$, respectively, and they are combined using the Assumptions C and D to give $F^{n:13}$. Moreover, as we have seen that $\bar{F}^{n:02}_{(1,0)} = \text{correct}$, $\bar{F}^{n:02}_{(1,1)} = \text{false}$ and $F^{n:12}_{(0,1)} = \emptyset$ and $F^{n:12}_{(1,1)} = \text{correct}$, hence there is no common setting (x_1, x_2) for which both $\bar{F}^{n:02}$ and $F^{n:12}$ are correct. Hence, there is no setting choice under which $F^{n:13}$ can be derived from $F^{n:12}$ and $\bar{F}^{n:02}$ as FR do. Agent F can always inherit both statements via the knowledge operator Eq. (C3), in a similar way to how you, the reader, is "inheriting" all the statements in this article when you read them. This, however, by itself poses little value due to the setting mismatch.

Alternatively, one can attempt a more direct derivation of the prediction associated with $F^{n:13}$, which tells us something about the outcome $w = \text{fail}$ based on the observation

of the outcome $z = 1/2$. This requires the setting $x_2 = 1$. Then, from F 's perspective under settings $(x_1, x_2) = (1, 1)$ and after obtaining measurement outcome $z = 1/2$, she would conclude that the probability of $w = \text{fail}$ is only

$$\begin{aligned}
 & P(w = \text{fail} | z = 1/2, (x_1, x_2) = (1, 1)) \\
 &= \sum_{r \in \{\text{heads}, \text{tails}\}} P(r | z = 1/2, (x_1, x_2) = (1, 1)) P(w = \text{fail} | r, z = 1/2, (x_1, x_2) = (1, 1)) \\
 &= P(w = \text{fail} | r = \text{tails} \& z = 1/2, (x_1, x_2) = (1, 1)) \\
 &= \frac{\text{tr} \left[\pi_{\text{fail}}^W \pi_{1/2}^F \pi_{\text{tails}}^{\bar{F}} \rho_{\text{Fi}} \pi_{\text{tails}}^{\bar{F}} \pi_{1/2}^F \right]}{\sum_{w \in \{\text{ok}, \text{fail}\}} \text{tr} \left[\pi_w^W \pi_{1/2}^F \pi_{\text{tails}}^{\bar{F}} \rho_{\text{Fi}} \pi_{\text{tails}}^{\bar{F}} \pi_{1/2}^F \right]} \\
 &= \frac{\text{tr} \left[\pi_{\text{fail}}^W \pi_{1/2}^F \pi_{\text{tails}}^{\bar{F}} \rho_{\text{Fi}} \pi_{\text{tails}}^{\bar{F}} \pi_{1/2}^F \right]}{\text{tr} \left[\pi_{1/2}^F \pi_{\text{tails}}^{\bar{F}} \rho_{\text{Fi}} \pi_{\text{tails}}^{\bar{F}} \right]} = \frac{1}{2} < 1, \tag{G16}
 \end{aligned}$$

where we have used Eq. (G15) and the last inequality follows from noting that if we are on branch 2.1.2, then the possibility of W measuring fail cannot be one, since he can get outcome ok, too (since, due to F 's measurement, the blue cancellation does not take place). Note also that the same probability is obtained under the settings $(x_1, x_2) = (0, 1)$.

The authors now proceed to reason from the perspective of the superagents. The first statement is $\bar{W}^{n:21}$: “System R is initialized to $|\text{init}\rangle_R$ at time $n:00$ ”. This statement is true for all settings, since it is a statement about the protocol, which all agents are assumed to know, so $\bar{W}_{(0,1)}^{n:21} = \bar{W}_{(1,0)}^{n:21} = \bar{W}_{(1,1)}^{n:21} = \text{correct}$. The authors then point out that *the state $U_{\bar{F}}|\text{init}, \phi_0, S_0\rangle$ [rhs of Eq. (D10)] is orthogonal to $|\bar{\text{ok}}\rangle_L |\downarrow\rangle_S$* . This is indeed correct, as can be readily seen from Fig. 6 by observing that $|\downarrow\rangle_S$ projects onto the superposition of branches 1.1.1 and 2.1.1 and that, for these branches, the purple $|\bar{\text{ok}}\rangle_L$ terms cancel each other out. The authors then state this in the form of an expectation value of a projector, namely, $\langle \text{init} | \pi_{(\bar{w}, z) \neq (\bar{\text{ok}}, -1/2)}^{n:00} | \text{init} \rangle = 1$. Here, $\pi_{(\bar{w}, z) \neq (\bar{\text{ok}}, -1/2)}^{n:00} = 1 - \pi_{(\bar{w}, z) = (\bar{\text{ok}}, -1/2)}^{n:00}$, where $\pi_{(\bar{w}, z) = (\bar{\text{ok}}, -1/2)}^{n:00}$ is a Heisenberg picture projector that would first transform $|\text{init}\rangle_R$ to $U_{\bar{F}}|\text{init}, \phi_0, S_0\rangle$ (through the appropriate isometry that appends $|\phi_0, S_0\rangle_{L \setminus S}$ and performs $U_{\bar{F}}$) and then projects onto the outcomes $(\bar{w}, z) = (\bar{\text{ok}}, -1/2)$, as $|\bar{\text{ok}}\rangle_L \langle \bar{\text{ok}}|_L \otimes |\downarrow\rangle_S \langle \downarrow|_S \otimes 1_L$; i.e., $\pi_{(\bar{w}, z) \neq (\bar{\text{ok}}, -1/2)}^{n:00}$ is the Heisenberg projector onto the complement of outcomes $(\bar{w}, z) = (\bar{\text{ok}}, -1/2)$. They then claim *agent $\bar{W}$, who uses Q , can, hence, be certain that $(\bar{w}, z) \neq (\bar{\text{ok}}, -1/2)$ and that this implies (when $\bar{w} = \bar{\text{ok}}$) the statement $\bar{W}^{n:22}$: “I am certain that F knows that $z = 1/2$ at time $n:11$ ”*. Now the authors are allowing $\bar{W}$ to take into account the measurement outcome of F in their reasoning but not the measurement of $\bar{F}$. In other words, they are using settings $(0, 1)$ and, thus, assigning

$$\begin{aligned}
 & P(z = 1/2 | \bar{w} = \bar{\text{ok}}, (x_1, x_2) = (0, 1)) \\
 &= \sum_{w \in \{\text{ok}, \text{fail}\}} P(z = 1/2, w | \bar{w} = \bar{\text{ok}}, (x_1, x_2) = (0, 1)) \\
 &= \sum_{w \in \{\text{ok}, \text{fail}\}} \frac{\text{tr} \left[\pi_w^W \pi_{\bar{\text{ok}}}^{\bar{W}} \pi_{1/2}^F \rho_{\text{Fi}} \pi_{1/2}^F \right]}{\sum_{z \in \{-1/2, 1/2\}} \sum_{w' \in \{\text{ok}, \text{fail}\}} \text{tr} \left[\pi_{w'}^W \pi_{\bar{\text{ok}}}^{\bar{W}} \pi_z^F \rho_{\text{Fi}} \pi_z^F \right]} \\
 &= \frac{\text{tr} \left[\pi_{\bar{\text{ok}}}^{\bar{W}} \pi_{1/2}^F \rho_{\text{Fi}} \pi_{1/2}^F \right]}{\sum_{z \in \{-1/2, 1/2\}} \text{tr} \left[\pi_{\bar{\text{ok}}}^{\bar{W}} \pi_z^F \rho_{\text{Fi}} \pi_z^F \right]} = 1, \tag{G17}
 \end{aligned}$$

where the last line follows from observing that in Fig. 6 we have that $\text{tr}[\pi_{\bar{\text{ok}}}^{\bar{W}} \pi_{-1/2}^F \rho_{\text{Fi}} \pi_{-1/2}^F] = 0$ and, thus, the rhs of the above is one, in accordance with what the authors claim. We, thus, have $\bar{W}_{(0,1)}^{n:22} = \text{correct}$.

While they are taking into account F 's measurement while reasoning, they are not taking into account $\bar{F}$'s measurement when reasoning. We can check that the statement $\bar{W}_{(1,1)}^{n:22} = \text{false}$ since the following probability, which takes into account both $\bar{F}$'s and F 's measurements, is strictly less than one:

$$\begin{aligned}
 & P(z = 1/2 | \bar{w} = \bar{\text{ok}}, (x_1, x_2) = (1, 1)) \\
 &= \sum_{\substack{w \in \{\text{ok}, \text{fail}\} \\ r \in \{\text{heads}, \text{tails}\}}} P(z = 1/2, w, r | \bar{w} = \bar{\text{ok}}, (x_1, x_2) = (1, 1)) \\
 &= \frac{\sum_{\substack{w \in \{\text{ok}, \text{fail}\} \\ r \in \{\text{heads}, \text{tails}\}}} \text{tr} \left[\pi_w^W \pi_{\bar{\text{ok}}}^{\bar{W}} \pi_{1/2}^F \pi_r^{\bar{F}} \rho_{\text{Fi}} \pi_r^{\bar{F}} \pi_{1/2}^F \right]}{\sum_{\substack{z \in \{-1/2, 1/2\} \\ w \in \{\text{ok}, \text{fail}\} \\ r \in \{\text{heads}, \text{tails}\}}} \text{tr} \left[\pi_w^W \pi_{\bar{\text{ok}}}^{\bar{W}} \pi_z^F \pi_r^{\bar{F}} \rho_{\text{Fi}} \pi_r^{\bar{F}} \pi_z^F \right]} \\
 &= \frac{\sum_{r \in \{\text{heads}, \text{tails}\}} \text{tr} \left[\pi_{\bar{\text{ok}}}^{\bar{W}} \pi_{1/2}^F \pi_r^{\bar{F}} \rho_{\text{Fi}} \pi_r^{\bar{F}} \pi_{1/2}^F \right]}{\sum_{\substack{z \in \{-1/2, 1/2\} \\ r \in \{\text{heads}, \text{tails}\}}} \text{tr} \left[\pi_{\bar{\text{ok}}}^{\bar{W}} \pi_z^F \pi_r^{\bar{F}} \rho_{\text{Fi}} \pi_r^{\bar{F}} \pi_z^F \right]} = \frac{1}{3} < 1. \tag{G18}
 \end{aligned}$$

Furthermore, $\bar{W}_{(1,0)}^{n:22} = \emptyset$, since the statement $\bar{W}^{n:22}$ is about F's measurement outcome. Before we move on, observe that Eqs. (G17) and (G18) take on different values and provides another example of collider bias. In particular, these two equations show that the probability of F's outcome does depend on $\bar{F}$'s setting x_1 given the knowledge that the postselection on $\bar{w}$ succeeded. This means that the probability $P(z = 1/2 | \bar{w} = \bar{ok})$ that FR consider is not well defined when the setting or the prior over the settings is not specified.

The authors then go on to claim ... *because agent $\bar{W}$ announces $\bar{w}$, agent W can be certain about $\bar{W}$'s knowledge, which justifies statement $W^{n:26}$, where $W^{n:26}$ is (assuming $\bar{W}$ announces $\bar{w} = \bar{ok}$ at time $n:21$) "I am certain that $\bar{W}$ knows that $\bar{w} = \bar{ok}$ at time $n:21$ ".* We agree that W can be sure of any correct announcement made

by $\bar{W}$, irrespective of their choice of (x_1, x_2) settings. Thus, $W_{(0,1)}^{n:26} = W_{(1,0)}^{n:26} = W_{(1,1)}^{n:26} = \text{correct}$.

Next, the authors make the claim ... *according to quantum mechanics, agent W can be certain that the outcome $(\bar{w}, w) = (\bar{ok}, ok)$ occurs after finitely many rounds.* They make this based on the fact that

$$\langle \psi_{Fi} | \pi_{(\bar{w}, w) = (\bar{ok}, ok)} | \psi_{Fi} \rangle = \frac{1}{12}, \quad (G19)$$

where $|\psi_{Fi}\rangle$ is the final state given on the rhs of Eq. (D11), $\pi_{(\bar{w}, w) = (\bar{ok}, ok)} = |\bar{ok}\rangle\langle\bar{ok}|_{\bar{L}} \otimes |ok\rangle\langle ok|_{\bar{L}}$. Note that this can also be seen as the case with $(x_1, x_2) = (0, 0)$, as we see later in Eq. (G23). The authors then claim $W^{n:00}$: "I am certain that there exists a round n in which the halting condition at time $n:40$ is satisfied." Noting that

$$\begin{aligned} P(\bar{w} = \bar{ok}, w = ok | (x_1, x_2) = (1, 1)) \\ &= \sum_{\substack{z \in \{-1/2, 1/2\} \\ r \in \{heads, tails\}}} \text{tr} \left[\pi_{ok}^W \pi_{ok}^{\bar{W}} \pi_z^F \pi_r^{\bar{F}} \rho_{Fi} \pi_r^{\bar{F}} \pi_z^F \right] \\ &= \sum_{\substack{z \in \{-1/2, 1/2\} \\ r \in \{heads, tails\}}} \left| \langle \bar{ok} | \langle ok | \pi_z^F U_F U_{Spin} \pi_r^{\bar{F}} U_{\bar{F}} | \text{init}, \phi_0, S_0 \rangle_{LS} | \phi_0 \rangle_L \right| = \frac{5}{6} \end{aligned} \quad (G20)$$

is positive, we conclude that $W_{(1,1)}^{n:00} = \text{correct}$. Similarly, we observe that

$$P(\bar{w} = \bar{ok}, w = ok | (x_1, x_2) = (0, 1)) = \sum_{z \in \{-1/2, 1/2\}} \text{tr} \left[\pi_{ok}^W \pi_{ok}^{\bar{W}} \pi_z^F \rho_{Fi} \pi_z^F \right] = \frac{1}{2} \quad (G21)$$

is positive; thus, $W_{(0,1)}^{n:00} = \text{correct}$. Likewise, there exists $r \in \{tails, heads\}$ such that

$$P(\bar{w} = \bar{ok}, w = ok | (x_1, x_2) = (1, 0)) = \sum_{r \in \{heads, tails\}} \text{tr} \left[\pi_{ok}^W \pi_{ok}^{\bar{W}} \pi_r^{\bar{F}} \rho_{Fi} \pi_r^{\bar{F}} \right] = \frac{1}{2} \quad (G22)$$

is positive; thus, $W_{(1,0)}^{n:00} = \text{correct}$. Finally,

$$P(\bar{w} = \bar{ok}, w = ok | (x_1, x_2) = (0, 0)) = \text{tr} \left[\pi_{ok}^W \pi_{ok}^{\bar{W}} \rho_{Fi} \right] = \frac{1}{12}; \quad (G23)$$

hence, $\bar{W}_{(0,0)}^{n:00} = \text{correct}$. Therefore, we conclude from Eqs. (G20)–(G23) that, while the statement $\bar{W}_{(x_1, x_2)}^{n:00}$ is correct for all settings, the value of the corresponding probability which FR calculate is not correct for all settings.

Next, the authors state *agent F may insert agent's $\bar{F}$'s statement $\bar{F}^{n:02}$ into $F^{n:12}$, obtaining statement $F^{n:13}$.* As we pointed out above, said statement holds only under certain settings. The authors then go on to say *by virtue of*

[assumption] *C, she [agent F] may then conclude that statement $F^{n:14}$ holds, too.* This statement is $F^{n:14}$: "I am certain that W will observe $w = \text{fail}$ at time $n:31$ ". This statement is merely the same as $F^{n:13}$ but now with the difference that F herself is certain that W will observe $w = \text{fail}$ at time $n:31$ rather than F being merely certain that $\bar{F}$ is certain. There are no settings F can select for which this statement is true, since

$$P(w = \text{fail} | z = 1/2, (x_1, x_2) = (x_1, 1)) < 1 \quad (G24)$$

for all settings $x_1 \in \{0, 1\}$. Moreover, F cannot inherit the statement "I am certain that W will observe $w = \text{fail}$ at time $n:31$ " via the transfer of knowledge operator, since this statement does not make reference to agent $\bar{F}$ and, thus,

TABLE II. Table 3 of Ref. [3] with the additional implicit assumptions needed to make each statement. Without the implicit assumptions, the implications drawn from Q in the original FR paper are summarized in the modal logic language in column 4. The corresponding explicit version of the same implications that take the additional implicit assumptions into account is given in the last column. While original implications can be combined using C and the distributive axiom to yield the further implications listed in Table 3 of the FR paper, the explicit version of these implications cannot be combined even using the standard rules of classical logic such as C and the distributive axiom. We, therefore, see that the paradox never arises even when using Q, C, and S and modeling agents unitarily as long as we are careful to use the explicit version of the implications. The variables x and y appearing in the last column are the settings appearing in the circuit in Fig. 9; these encode the additional implicit assumptions.

Agent	Assumed observation	Statement inferred via Q	Original implication obtained from Q	Additional implicit assumption	Explicit implication obtained from Q
$\bar{F}$	$r = \text{tails}$ at $n:01$	$\bar{F}^{n:02}$: I am certain that W will observe $w = \text{fail}$ at $n:31$	$K_{\bar{F}}(r = \text{tails} \Rightarrow w = \text{fail})$	$\bar{F}$'s outcome is reasoned about and F's lab is modeled as a closed quantum system	$K_{\bar{F}}((x, y) = (1, 0) \wedge r = \text{tails} \Rightarrow w = \text{fail})$
F	$z = +\frac{1}{2}$ at $n:11$	$F^{n:12}$: I am certain that $\bar{F}$ knows that $r = \text{tails}$ at $n:01$	$K_F(z = +\frac{1}{2} \Rightarrow K_{\bar{F}}(r = \text{tails}))$	$\bar{F}$'s and F's outcomes are both reasoned about	$K_F((x, y) = (1, 1) \wedge z = +\frac{1}{2} \Rightarrow K_{\bar{F}}(r = \text{tails}))$
$\bar{W}$	$\bar{w} = \bar{\text{ok}}$ at $n:21$	$\bar{W}^{n:22}$: I am certain that F knows $z = +\frac{1}{2}$ at $n:11$	$K_{\bar{W}}(\bar{w} = \bar{\text{ok}} \Rightarrow K_F(z = +\frac{1}{2}))$	F's outcome is reasoned about and $\bar{F}$'s lab is modeled as a closed quantum system	$K_{\bar{W}}((x, y) = (0, 1) \wedge \bar{w} = \bar{\text{ok}} \Rightarrow K_F(z = +\frac{1}{2}))$
W	Announcement by $\bar{W}$ that $\bar{w} = \bar{\text{ok}}$ at $n:21$	$W^{n:26}$: I am certain that $\bar{W}$ knows that $\bar{w} = \bar{\text{ok}}$ at $n:21$	$K_W K_{\bar{W}}(\bar{w} = \bar{\text{ok}})$	(None)	$K_W K_{\bar{W}}(\bar{w} = \bar{\text{ok}})$

it would require F *himself* to use settings $(x_1, x_2) = (1, 0)$ just after obtaining the measurement outcome $z = \frac{1}{2}$; yet the latter requires setting $x_2 = 1$ (this is distinct to the case of $F^{n:13}$).

The remaining statements made by W and $\bar{W}$ are derived from the previous statements under the assumptions Q, U, C, D, and S. However, they do so disregarding the setting parameters and, hence, reaching their apparent contradiction.

In Table II, we compare the original statements of the FR paper used in deriving the apparent contradiction and the explicit version of those statements obtained within our framework. This is analogous to Table I of the entanglement case, and it can be immediately seen that, while the original statements (which ignore setting information) can be combined to yield the apparent contradiction, the explicit statements (which specify the setting choice) cannot be combined in this manner even using the standard rules of classical logic. These explicit statements can also be derived directly within the augmented circuit of the EWFS at hand, which is the prepare-and-measure version of the FR scenario. This augmented circuit is illustrated in Fig. 9, and an equivalent version of this circuit (that makes the mapping to the entanglement formulation of the scenario more evident) is given in Fig. 10.

In summary, in this section, we have shown that each of the individual statements that FR use to derive a

contradiction (see Table II) hold under some choice of settings in our framework (note this also follows from Appendix J 2 of the general framework). However, each of these statements requires a different setting choice and can no longer be combined using the FR assumptions to yield a paradox.

3. Setting dependence in FR's scenario

In Appendix G 1, we mapped each FR probability [$P_{\text{conv}}(b = 1|u = w = \text{ok}) = 1$, $P_{\text{conv}}(a = 1|b = 1) = 1$, and $P_{\text{conv}}(w = \text{fail}|a = 1) = 1$] to a unique setting choice $(x_1, x_2) \in \{(1, 0), (1, 1), (0, 1)\}$ that reproduces the same probability in our framework. It is illustrative to go beyond this analysis and consider other possible setting choices we can assign to each statement. Then, we find that, in contrast to the above-mentioned settings which do reproduce the exact probabilities of the FR paper, alternate setting choices no longer give the same probabilities and, therefore, do not yield the desired logical statements needed in FR's reasoning. For instance, we showed in Appendix G 1 that $P_{\text{conv}}(b = 1|u = w = \text{ok}) = P(b = 1|u = w = \text{ok}, (x_1, x_2) = (0, 1)) = 1$. We can alternatively consider $P(b = 1|u = w = \text{ok}, (x_1, x_2) = (1, 1))$. Note that we cannot consider any setting choice with $x_2 = 0$ here, because the prediction involves Bob's outcome $b = 1$ and, therefore, must involve a nontrivial projector on Bob's side in its evaluation:

$$\begin{aligned}
P(b = 1, u = w = \text{ok} | (x_1, x_2) = (1, 1)) &= \frac{1}{12} \\
&= \sum_{a \in \{0,1\}} \left| \langle \text{ok} |_{\text{RA}} \otimes \langle \text{ok} |_{\text{SB}} \left(\pi_{x_1=1,a}^A \otimes \pi_{x_2=1,b=1}^B \right) | \psi^{t=2} \rangle \right|^2 \\
&= \sum_{a \in \{0,1\}} \left| \langle \text{ok} |_{\text{RA}} \otimes \langle \text{ok} |_{\text{SB}} \left(|aa\rangle \langle aa|_{\text{RA}} \otimes |11\rangle \langle 11|_{\text{SB}} \right) | \psi^{t=2} \rangle \right|^2, \quad (\text{G25})
\end{aligned}$$

$$\begin{aligned}
P(u = w = \text{ok} | (x_1, x_2) = (1, 1)) &= \frac{1}{4} \\
&= \sum_{a,b \in \{0,1\}} \left| \langle \text{ok} |_{\text{RA}} \otimes \langle \text{ok} |_{\text{SB}} \left(\pi_{x_1=1,a}^A \otimes \pi_{x_2=1,b}^B \right) | \psi^{t=2} \rangle \right|^2 \\
&= \sum_{a,b \in \{0,1\}} \left| \langle \text{ok} |_{\text{RA}} \otimes \langle \text{ok} |_{\text{SB}} \left(|aa\rangle \langle aa|_{\text{RA}} \otimes |bb\rangle \langle bb|_{\text{SB}} \right) | \psi^{t=2} \rangle \right|^2. \quad (\text{G26})
\end{aligned}$$

Putting this together using the rule for conditional probabilities, we have

$$P(b = 1 | u = w = \text{ok}, (x_1, x_2) = (1, 1)) = \frac{1}{3} \neq 1. \quad (\text{G27})$$

The FR probability $P_{\text{conv}}(a = 1 | b = 1) = 1$ corresponds to $P(a = 1 | b = 1, (x_1, x_2) = (1, 1)) = 1$ in our framework, but no alternative setting choices are possible here, as the prediction refers to both Alice and Bob's classical outcomes. Finally, the FR probability $P_{\text{conv}}(w = \text{fail} | a = 1) = 1$ corresponds to $P(w = \text{fail} | a = 1, (x_1, x_2) = (1, 0)) = 1$ in our framework, and we consider the alternative setting choice $(x_1, x_2) = (1, 1)$ in this case (again, $x_1 = 0$ is not a possible setting choice here, as the prediction involves Alice's classical outcome). The corresponding prediction $P(w = \text{fail} | a = 1, (x_1, x_2) = (1, 1))$ is calculated below and is also not equal to 1:

$$\begin{aligned}
P(a = 1, w = \text{fail} | (x_1, x_2) = (1, 1)) &= \frac{1}{3} \\
&= \sum_{b \in \{0,1\}} \left| 1_{\text{RA}} \otimes \langle \text{fail} |_{\text{SB}} \left(\pi_{x_1=1,a=1}^A \otimes \pi_{x_2=1,b}^B \right) | \psi^{t=2} \rangle \right|^2 \\
&= \sum_{b \in \{0,1\}} \left| 1_{\text{RA}} \otimes \langle \text{fail} |_{\text{SB}} \left(|11\rangle \langle 11|_{\text{RA}} \otimes |bb\rangle \langle bb|_{\text{SB}} \right) | \psi^{t=2} \rangle \right|^2, \quad (\text{G28})
\end{aligned}$$

$$\begin{aligned}
P(a = 1 | (x_1, x_2) = (1, 1)) &= \frac{2}{3} \\
&= \sum_{b \in \{0,1\}} \left| \left(\pi_{x_1=1,a=1}^A \otimes \pi_{x_2=1,b}^B \right) | \psi^{t=2} \rangle \right|^2 \\
&= \sum_{b \in \{0,1\}} \left| \left(|11\rangle \langle 11|_{\text{RA}} \otimes |bb\rangle \langle bb|_{\text{SB}} \right) | \psi^{t=2} \rangle \right|^2. \quad (\text{G29})
\end{aligned}$$

This gives us

$$P(w = \text{fail} | a = 1, (x_1, x_2) = (1, 1)) = \frac{1}{2} \neq 1. \quad (\text{G30})$$

In summary, we have found that $P(b = 1 | u = w = \text{ok}, (x_1, x_2) = (0, 1)) \neq P(b = 1 | u = w = \text{ok}, (x_1, x_2) = (1, 1))$ and $P(w = \text{fail} | a = 1, (x_1, x_2) = (1, 0)) \neq P(w = \text{fail} | a = 1, (x_1, x_2) = (1, 1))$. That is, the predictions do depend on the setting. This highlights that a core reason for the FR paradox in this scenario is that the FR reasoning involves ignoring the conditioning on setting values in a scenario where the predictions do depend on these value. This illustrates our general result of Corollary 2, which shows that, in any EWFS, any potential inconsistencies can arise only in this manner: ignoring the conditioning on settings, in predictions which are setting dependent.

Reformulating this insight in a more framework-independent manner, this means that the choice of Heisenberg cuts under which we evaluate the predictions for different sets of outcomes does matter, and the probabilities do depend on this choice (captured by the settings in our framework). Our framework, which explicitly takes into account the Heisenberg cut, provides a natural resolution to any such paradox without forsaking unitary quantum theory or classical logic.

APPENDIX H: CLASSICAL EXAMPLE REPRODUCING CERTAIN FEATURES OF THE FR CORRELATIONS

We noted earlier that the FR correlations are an example of collider bias, or, in this case, they exhibit signaling under postselection. Furthermore, we have also seen that the assumption **I** is necessary for reproducing the apparent paradox. In this section, we provide an example of a classical protocol that reproduces these features of the FR correlations. We note, however, that the settings in our example do not have the same physical meaning as the settings of the FR augmented circuit (which relate to the choice of Heisenberg cut). Nevertheless, the example gives an intuition for our resolution. We, thus, refer to the **I** applied to the settings of this classical example as **I_C** to distinguish it from the version of the assumption applied to the settings of our augmented circuit that model Heisenberg cuts.

The augmented circuit in Fig. 5 of the entanglement version of the FR protocol can be represented as shown in Fig. 11(a). Consider now the circuit in Fig. 11(b), which has the same configuration as the circuit in Fig. 11(a) for the FR protocol, and consider the following classical

operations in place of the boxes $\mathcal{A}$, $\mathcal{B}$, $\mathcal{U}$, and $\mathcal{W}$. The initial state is simply a uniformly distributed binary random variable Λ . The operation $\mathcal{A}$ of Alice takes the given value $\lambda \in \{0, 1\}$ of Λ together with the binary setting x_1 and generates the outcome $a \in \{0, 1\}$ by taking the XOR of the two $a = \lambda \oplus x_1$ and forwards a to Ursula's operation $\mathcal{U}$. $\mathcal{U}$ takes a from Alice's operation, internally generates a uniformly distributed bit $k_U \in \{0, 1\}$, and outputs $u = a \cdot k_U$, where $\cdot$ denotes the logical AND. Bob's side is identical; his operation $\mathcal{B}$ takes a value of Λ along with x_2 and outputs $b = \lambda \oplus x_2$ while forwarding a copy of the classical bit b to Wigner's operation $\mathcal{W}$. $\mathcal{W}$ takes b from Bob, generates a uniform bit k_W internally, and outputs $w = b \cdot k_W$.

Given this simple model, it is easy to verify that $P(a, b|x_1, x_2)$ is nonsignaling, i.e., $P(a|x_1, x_2) = P(a|x_1)$ and $P(b|x_1, x_2) = P(b|x_2)$. However, $P(a, b|x_1, x_2, u)$ allows signaling from Alice to Bob, i.e., $P(b|x_1, x_2, u) \neq P(b|x_2, u)$, and $P(a, b|x_1, x_2, w)$ allows signaling from Bob to Alice, i.e., $P(a|x_1, x_2, w) \neq P(a|x_1, w)$. This is because, for instance, conditioned on the knowledge that $u = 1$ (say), we know that, since $u = a \cdot k_U$, $a = k_U = 1$. Furthermore, since $a = \lambda \oplus x_1 = 1$ and $b = \lambda \oplus x_2$, we know that $b = x_1 \oplus x_2 \oplus 1$ and Bob's outcome b clearly depends on Alice's setting x_1 now. This means that when reasoning given the knowledge about some postselections on u and w , even in classical probability theory, the agents must be careful to account for any new correlations that this may introduce between their settings; else they may end up making wrong conclusions.

Interestingly, in Ref. [16], Healey pointed out that whether or not Bob's state is collapsed influences the

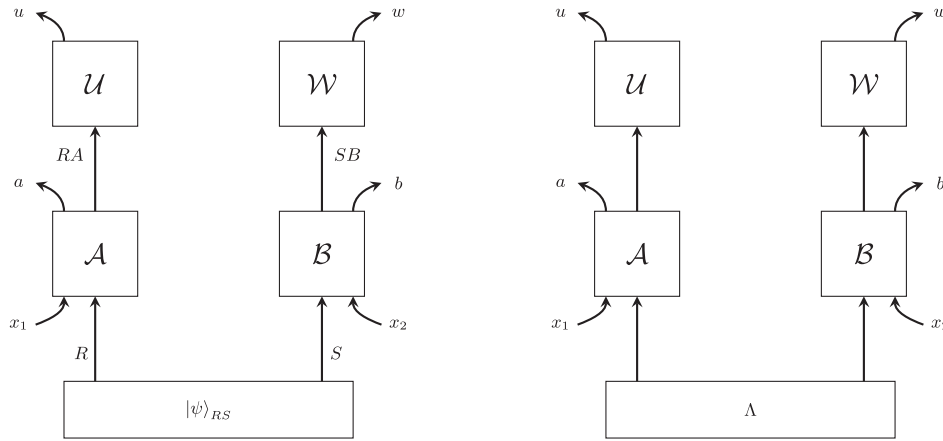


FIG. 11. Left: a concise representation of the augmented circuit (Fig. 5) corresponding to the entanglement version of the FR protocol. Here, the boxes $\mathcal{A}$ and $\mathcal{B}$ correspond to Alice and Bob's measurements modeled as unitaries $M_{\text{unitary}}^{\mathcal{A}}$ and $M_{\text{unitary}}^{\mathcal{B}}$ followed by the setting-dependent projectors $\{\pi_{x_1}^{\mathcal{A}}\}$ and $\{\pi_{x_2}^{\mathcal{B}}\}$, respectively. The boxes $\mathcal{U}$ and $\mathcal{W}$ model to ok, fail basis measurements of Ursula and Wigner on the joint systems RA and SB producing outcomes u and w . Right: a classical circuit with the same configuration as the left-hand side circuit of the FR protocol. Here, Λ , a , b , u , w , x_1 , and x_2 are all associated with classical, binary variables, and the operations $\mathcal{A}$, $\mathcal{B}$, $\mathcal{U}$, and $\mathcal{W}$ and initial distribution over Λ are described in the text. Here as well, as in the FR scenario, x_1 and b , although initially independent (due to no signaling), become correlated when conditioning on the future outcome w , and similarly x_2 and a become correlated under postselection on u .

probabilities of Alice's outcome in the FR scenario due to the postselection. Denying this dependence is what Ref. [16] refers to as the assumption of *intervention insensitivity*. Healey argues that FR implicitly assume intervention insensitivity in order to obtain the logical contradiction. While Healey's intervention sensitivity may seem similar in spirit to the setting dependence in our framework, Ref. [16] appears to regard intervention sensitivity as a special nonlocal feature of quantum theory. In the FR case, the setting dependence takes the particular form of enabling signaling under postselection that we have explained here. However, the present classical example shows that this is not a special feature of the quantum correlations of the FR scenario but rather a common feature that arises in classical probability theory and causal inference [48], where it often goes by the name of *collider bias*.

In the FR protocol, this feature is not immediately apparent, as there are no settings in the actual protocol. However, as we have argued in the main text, these settings are implicit in the reasoning steps, and, once made explicit, we see that there is indeed signaling under postselection also in the FR scenario. Note that our general resolution and results do not depend on this aspect of signaling under postselection or collider bias, although this is a feature of FR correlations. The general results show that imposing the **I** assumption in a scenario where it fails is necessary for recovering FR-type apparent paradoxes (independently of this feature), and we have seen that **I** does fail in the FR scenario.

This is also the case in our classical example; we can show that imposing **I_C** also leads to an apparent paradox here, as the predictions do depend on the settings (the scenario violates **I_C**). For instance, take $u = 1$ and $x_2 = 1$; then, we have $a = k_u = 1$ and, hence, $\Lambda = x_1 \oplus 1$, which gives $b = x_1 \oplus x_2 \oplus 1 = x_1$. Therefore, we have $P(b = 0|x_1 = 0, u = 1, x_2 = 1) = 1$ and $P(b = 1|x_1 = 1, u = 1, x_2 = 1) = 1$. If we ignore the conditioning on the settings, as allowed by **I_C**, we apparently have the paradoxical probability assignments $P(b = 0|u = 1) = 1$ and $P(b = 1|u = 1) = 1$ that correspond to b deterministically being 0 and 1 (both with certainty) whenever u is 1, which is an apparent violation of **S** as in the FR scenario. However, we see that there is no real paradox once we correctly take into account all the settings and account for the correlations of the outcomes with the settings.

While the property of signaling under postselection holds in this simple classical circuit as it does in the quantum FR scenario, the exact correlations of the FR scenario are Bell nonlocal (or, more generally, contextual) and cannot be reproduced using classical resources alone, once we fix the above circuit configuration where there is no information exchange between Alice and Bob's sides (except the shared initial state). In order to reproduce the FR correlations classically, one must necessarily modify

the circuit configuration to allow for additional connections such as allow the settings x_1 and x_2 to depend on the state preparation or allow for causal connections between the $\mathcal{A}$, $\mathcal{U}$ and $\mathcal{B}$, $\mathcal{W}$ operations, analogous to superdeterministic or nonlocal hidden variable explanations of Bell correlations (see, for instance, Ref. [68]). The correspondence between the FR argument and Hardy's logical argument for contextuality of quantum correlations is discussed in the next section.

APPENDIX I: RELATION TO PREVIOUS WORKS: A MORE UNIFIED PICTURE

There are several previous works discussing and analyzing the FR apparent paradox, which can be categorized into three broad groups. The first category includes works (e.g., Refs. [40,69–72]) which provide fundamental reasons for rejecting one of the assumptions of the FR no-go theorem. This is a natural response to any no-go theorem, as it necessitates identifying which assumptions to discard.

The second category consists of papers such as Refs. [4,12,73,74] which (possibly after identifying implicit assumptions beyond **Q**, **C**, and **S**) conclude that there is indeed a paradox in the FR scenario. Some of these papers suggest that FR's result raises deeper problems for foundations or interpretations of quantum theory (e.g., Ref. [73]), while others (e.g., Ref. [12]) suggest possible resolutions by proposing additional reasoning rules.

The third category includes works such as Refs. [13,15–18,31,75,76] that question the correctness of the theorem. They argue that the reasoning in the FR paper includes implicit assumptions, one or more of which are invalid in quantum theory, and conclude that there is no real paradox between quantum theory and logic as FR claim [77].

These responses are often discussed independently and can be interpretation specific. Here, we provide a more unified picture of these arguments within our framework, also discussing how different interpretations of quantum mechanics could apply our framework to resolve the apparent paradox in any EWFS.

1. Previous works rejecting one of FR's assumptions

Here, we consider previous arguments which fundamentally reject one of the assumptions **Q**, **U**, **C**, **D**, or **S** in the FR no-go theorem [3,4]. We discuss these cases in detail below, relating them to different interpretations of quantum mechanics and demonstrating how these arguments play out within our framework.

a. Rejecting **Q**

In FR's original paper, the authors interpret the violation of **Q** as a violation of unitary quantum theory. Following Ref. [4], we have separated these two assumptions **Q** and **U** explicitly, such that the **Q** assumption refers to the validity

of the Born rule (or a weaker possibilistic version thereof), which can be independently violated without giving up $\mathbf{U}$.

Rejecting $\mathbf{Q}$ means at least one prediction of the FR scenario [e.g., Eq. (12) for the entanglement version] does not comply with the Born rule. Not all predictions can be simultaneously tested in a single experiment, leaving room for observable compliance with the Born rule without satisfying $\mathbf{Q}$. As noted in Ref. [78], certain versions of Bohmian mechanics, a nonlocal hidden variable interpretation, violate $\mathbf{Q}$ in this manner, by deviating from the Born rule for inaccessible predictions.

In our framework, predictions in an EWFS are given by $P(\vec{a}_j|\vec{a}_l, k)$ (Definition 2), where $\vec{a}_j$ and $\vec{a}_l$ are sets of measurement outcomes and k are a set of parameters describing the scenario. In hidden variable interpretations, in addition to quantum states, channels, and measurements (Definition 1), k can include descriptions of the hidden variables. $P(\vec{a}_j|\vec{a}_l, k)$ may be computed using this information, not necessarily following the Born rule.

Such predictions can violate assumption $\mathbf{Q}$ formalized in our work, and they would not correspond to the setting-conditioned predictions of the augmented circuit (Definition 8) which are derived using the Born rule. However, the settings are still meaningful in such interpretations; they provide different descriptions of the quantum measurement channel that the hidden variables must “emulate.” It is to be noted that, within such theories, settings cannot be interpreted as choices of Heisenberg cuts, since classical theories do not have a nontrivial notion of such a cut. But such a fully classical theory could “emulate” a unitary measurement channel or setting $x_i = 0$ through nonlocal hidden mechanisms.

Moreover, by applying the general premise of our reasoning rules discussed in Sec. VII A, FR-type paradoxes can also be avoided in any EWFS within such interpretations as well. This would involve using settings to clearly specify the measurement channels being assumed, as well as potentially additional parameters in k and conditioning on these variables in the reasoning (even when the probabilities do not arise through the Born rule).

b. Rejecting $\mathbf{U}$

The $\mathbf{U}$ assumption, first noted in Ref. [4], involves rejecting the idea that agents’ labs evolve as unitary closed quantum systems. Interpretations involving objective mechanisms for “wave function collapse” reject this, where unitary quantum theory breaks down at certain macroscopic scales. Such interpretations also violate the $\mathbf{U}$ assumption in our work. In this view, one would assign setting $x_i = 1$ for all measurements, assuming they involve systems more macroscopic than the scale at which the collapse occurs. Whether such models always avoid FR-type paradoxes depends on whether the predicted objective collapse scale is smaller than the scale at which a quantum system can be regarded as an “agent,” which remains an open question.

In our formalization of the $\mathbf{U}$ assumption as $\mathbf{U}$, we have made explicit an additional aspect that is often implicit in previous works, relating to quantum control over labs of other agents. Interpretations compatible with unitary quantum theory can still violate $\mathbf{U}$ by arguing that full quantum control over an agent’s lab is impractical. For instance, in decoherence-based interpretations, measurements are always associated with an inaccessible environment that decoheres the quantum superposition. To regard each $\mathcal{M}^{A_i}$ in our EWFS definition (Definition 1) as a “measurement” according to this viewpoint, one must restrict to standard quantum scenarios (Definition 13), where one agent cannot have complete quantum control over the labs of other agents. As shown by Corollary 1, in these scenarios, a joint distribution for all (nontrivial) measurement outcomes can be assigned, which would be equivalent to the distribution obtained by choosing $x_i = 1$ for all settings.

Notably, a recent work [40] makes an argument for objective decoherence based on the concept of premeasurements. The notion of a premeasurement coincides with the case of setting 0 in our framework, i.e., the modeling of the measurement purely as a unitary evolution. They argue that premeasurements cannot produce outcomes consistently in quantum theory, and, to produce an outcome, one requires an irreversible evolution. This aspect is captured within our framework in the fact that no nontrivial measurement outcome can be assigned to a measurement modeled with setting 0; nontrivial outcomes require setting 1 (applying the projection postulate). The authors of Ref. [40] then suggest that this observation resolves the FR apparent paradox, as FR’s reasoning assigns outcome values to premeasurements. However, Ref. [40] imposes an objective distinction between premeasurements and measurements, considering the operations of the agents Alice and Bob only as premeasurements (setting 0) and those of the superagents Ursula and Wigner as measurements (setting 1). Moreover, the criterion for objectively fixing this choice in general EWFSs is not explicitly discussed in prior works.

The resolution proposed here is significantly more general. It applies to all EWFSs, allowing for generally subjective setting assignments (given by an explicit rule) while consistently accounting for relational interpretations as well.

c. Rejecting $\mathbf{C}$ or $\mathbf{D}$

The assumption $\mathbf{C}$ allows agents to inherit each other’s knowledge. FR suggested that relational approaches, such as relational quantum mechanics [79] and QBism [80,81], reject this assumption by allowing different subjective perspectives. However, since the original $\mathbf{C}$ assumption left unclear the formal modeling of agents’ knowledge and choices of Heisenberg cuts, we have proposed a formalization of this assumption in our framework, as $\mathbf{C}$.

In subjective and relational approaches to EWFSs, agents can have different choices of Heisenberg cuts.

Superagents model agents as “inside the cut,” assuming unitary evolution of their labs, while agents model themselves as “outside the cut,” associating classical outcomes to their measurements. This is captured in our framework through settings; our reasoning rule (see Appendix J 2 and Sec. VII A) allows the choices of these settings to be subjective and agent dependent in a precise manner.

Despite this relational aspect and allowing dynamic updates of agents’ knowledge (based on new observations), our framework still allows agents to freely inherit each other’s knowledge, as it satisfies the assumption C. Even when agents use different settings, if assumptions about setting choices are explicitly stated, C and other logical rules, such as the distributive axiom D and transitivity, are always satisfied as shown in Corollary 2. Ignoring these setting choices leads to apparent breakdowns of these axioms, but, as argued in Sec. VII A, similar logical breakdowns can occur in classical scenarios if assumptions about channels used in reasoning are ignored.

There also exist proposals, such as Ref. [82], which reject classical logical axioms in light of FR’s work. Our results highlight that such rejection is unnecessary for preserving the consistency of quantum theory or relational interpretations.

d. Rejecting S

Previous discussions [3,4,11] suggest that many-worlds interpretations [83] would tend to reject S, because an outcome a_i could be 0 in one “branch” of the wave function and 1 in another. Treating a_i as a quantum object associated with a quantum state $|a_i\rangle$, the violation of S seems less paradoxical, since the quantum state not being $|0\rangle$ does not imply it is $|1\rangle$ but could be another nonorthogonal state [18]. Violation of S is paradoxical only when a_i is regarded as a classical random variable.

In our framework, there is no ambiguity between classical and quantum objects. Even with subjective points of view, agents’ measurement outcomes are classical variables, with the trivial value $a_i = \perp$ when the measurement is unitary. Therefore, rejecting S within our framework would indeed constitute a paradox, leading to an invalid probability distribution $P(a_i)$. However, we have shown this does not happen in our framework.

Thus, many-worlds interpretations would not reject S in our framework. We do not believe any previous proposal or interpretation of quantum theory is pathological enough to fail this assumption. Moreover, it is important to note that selecting setting $x_i = 1$ when computing probabilities of a_i does not conflict with many-worlds interpretations where the Universe undergoes unitary evolution. The projectors applied for the $x_i = 1$ case need not be interpreted as a physical projection or wave function collapse but can be seen as a mathematical tool necessary for computing probabilities through the Born rule. Alternatively, this can also be understood akin to classical probability theory,

as conditioning on an agents’ knowledge of seeing a particular outcome $a_i = a_i$ in a particular “branch” of a many-worlds wave function, even when other branches can exist, from the perspective of other agents. It would be interesting to extend such ideas to formally define a perspectival version of the many-worlds interpretation, from an operational approach.

2. Previous suggestions for consistent reasoning

Another category of previous works is those that propose ways around FR’s apparent paradox, either through conceptual discussions or by suggesting additional reasoning rules.

We have shown that, if the FR scenario and its conventional quantum predictions are equivalently described using our augmented circuit, there is no paradox, eliminating the need for additional reasoning rules [84]. Our formal version of each FR assumption is satisfied without contradictions. Nevertheless, it is insightful to interpret previous works’ proposed reasoning rules.

a. Adding context labels

At a more pedagogical level, Ref. [4] suggests an additional reasoning rule that could be incorporated to avoid the paradox, which involves tagging all statements made by agents with certain “context labels” and combining only statements with matching labels. However, no particular model for these labels or a rigorous formalization of the rule was proposed there, and the question of efficiency and causal consistency (namely, regarding the number of contexts to be checked at each step and whether one needs to account for the contexts of future measurements) was left open. The settings of our framework can be interpreted as a concrete instantiation of these abstract context labels, but we have seen that no additional reasoning rules beyond quantum theory and classical logic are required once the settings are explicitly specified.

Importantly, our framework does not impose that only statements derived from the same setting labels can be combined. In fact, we have shown that in certain situations the statements become independent of the certain setting labels and the choice of those setting labels no longer affects the validity of the statement (see Appendix J 2 and Sec. VI B). We have also addressed the efficiency and causality issues.

b. Parsing rule for measurements

Another recent work [12] proposes a parsing rule for quantum theory to determine when a unitary or isometry can be considered a measurement with a classical outcome. This rule deems an operation a measurement if no future noncommuting operations act on the memory system, ensuring consistency in the FR scenario by limiting agents’ reasoning to such operations. This idea also aligns with the

principle of superpositional solipsism in Ref. [74]. Unlike our framework, which uses settings to describe measurements, Ref. [12] maintains ambiguity in measurement modeling, but the additional parsing rule clarifies when the ambiguity can be safely ignored in agents' reasoning.

If an operation $\mathcal{M}^{A_i}$ in the EWFS description is deemed a measurement according to this parsing rule, then it would imply that there are no superagents to the agent A_i in the scenario (also according to our definition of the nonsuperagent structure). As shown in Theorem F1, A_i 's setting can then be safely ignored in the predictions, allowing all agents to reason about A_i 's outcome without inconsistencies in that scenario. However, we note that our non-superagent structure is formalized without reference to commutativity of operations. In particular, a noncommuting operation acting solely on the memory M_i might fail the parsing rule of Ref. [12] but would still allow A_i 's setting to be safely ignored in our framework.

Moreover, our framework does not restrict the ability to reason about A_i 's outcome based on whether there are superagents to A_i (i.e., someone “Hadamarding” A_i 's brain) in the scenario. As we have seen, each of FR's statements can be reproduced in our formalism. This is in contrast to Ref. [12] and another work [14], which suggest that certain statements of FR that refer to outcomes of agents' whose brains will later be Hadamarded should not be allowed. While such additional rules or restrictions are sufficient to ensure logical consistency, the adherence to causality principles and efficiency of programming remain open question. Our work shows that such additional rules are not necessary for the purpose of ensuring consistency in any EWFS and that logical, causal consistency and efficiency of reasoning are possible with weaker restrictions on the reasoning, though it may be of interest to impose such rules based on other physical considerations or interpretations.

c. Consistent histories interpretation

Another notable approach is the consistent histories (CH) interpretation of quantum theory [38,39]. In this approach, one specifies a set of possible histories for a given scenario and a consistency criterion that tells us when a set of histories is consistent, allowing probabilities to be assigned to such consistent sets. The approach has been applied to explain a number of quantum paradoxes arising in standard (non-Wigner's-friend-like) quantum scenarios, such as contextuality paradoxes, pre- and postselection paradoxes, etc. In the context of Wigner's friend scenarios, Ref. [13] shows that the reasoning used in FR's derivation of the paradox requires computing probabilities in an inconsistent family of histories, and the authors then argue that such reasoning is, therefore, not valid in quantum theory.

At a high level, this bears resemblance to the general result of Corollary 3, where we demonstrated that any EWF

paradox is rooted in computing predictions across distinct setting choices and combining such statements while ignoring the setting labels (even though the predictions depend on these labels). However, our approach is distinct from the CH interpretation in key ways and arguably offers a simpler and more minimal resolution to EWFS paradoxes.

The CH approach requires considering commutation relations between families of projectors to determine consistency, while our settings are formalized without reference to commutation relations and do not involve such additional rules to ensure consistency. We have seen that conditioning on settings that model a measurement corresponds to conditioning on the choice of channel used in computing a probability; no additional rule is required to forbid combinations of statements made under different settings (as is needed in the CH approach to avoid combining inconsistent histories).

Moreover, our setting independence results highlight that, in certain cases, classical probability theory and classical logic ensure that even statements made under different settings can be consistently combined, as those statements are independent of the setting choice. While noncommuting projectors central to the CH approach are a nonclassical aspect, we have discussed how our resolution of the paradox shares strong similarities with how analogous multiagent inconsistencies are resolved in purely classical theories (see also the previous paragraph on parsing rules for further discussion on the link between noncommutativity and our resolution).

Moreover, while CH's solution is sufficient to avoid logical inconsistencies, it does not provide an explicit reasoning rule for how to select the set of histories to be used when reasoning about predictions or agents' knowledge in an EWFS [11]. In particular, Ref. [11] noted that, if unitary quantum theory were universally valid, the perspectives and predictions of different agents would indeed correspond to different histories, and, even in a single experiment such as the FR scenario, there is no single objective history of events that is realized.

A natural question that arises is whether there is a unified framework with a concrete set of rules to construct it where all these perspectives and predictions of an EWFS can be consistently incorporated while recovering the predictions of real-world quantum experiments performed so far. Here, we have demonstrated that all EWFSs in quantum theory can be completely described within a single consistent quantum circuit framework that is capable of resolving general EWF paradoxes. We also provided an explicit rule for selecting the settings (modeling the Heisenberg cuts) in accordance with universal validity of unitary quantum theory and without assuming the existence of an objective notion of observed measurement events; the observations and predictions in our formalism are relative to a choice of such settings.

3. Previous works discussing the validity of FR's claim

A third category of papers are those which question the validity of FR's theorem, due to additional implicit assumptions (other than Q, U, C, D, and S discussed in Refs. [3,4]) which are violated [13,15–18,76]. Some specific examples include Aaronson's blog post that refers to an additional "unformalized" assumption of FR, Healey's assumption of intervention insensitivity [16], and the assumption regarding collapse or no collapse that Araujo points out [17]. To quote Aaronson: But I reject an assumption that Frauchiger and Renner never formalise. That assumption is, basically: "it makes sense to chain together statements that involve superposed agents measuring each other's brains in different incompatible bases, as if the statements still referred to a world where these measurements weren't being done."

In our work, we have concretely shown that an additional assumption **I** (setting independence) is violated in the FR scenario but must be necessarily imposed for reproducing the apparent paradox. We have seen that the assumptions **Q**, **U**, **C**, **D**, and **S** about the validity of quantum theory and classical logic are always consistent in any EWFS in our formalism. In our understanding, the assumption **I** formally embodies the spirit of the additional implicit assumptions noted in the above examples of previous works. All these previous works argue that the respective assumption is necessary to reproduce the apparent paradox of FR but suggest that the assumption fails in the FR scenario, due to the incompatible measurements and/or (Bell) nonlocality of the correlations involved.

While the assumption **I** of our framework is formulated in a much more general manner and is *a priori* independent of quantum features such as incompatible measurements and nonlocality, when applied to the FR scenario, it captures the features of these previously noted assumptions. Aaronson's assumption is reflected in our framework by noting that the four possible settings $(x_1, x_2) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$ for the FR protocol are in one-to-one correspondence with the four possible measurement contexts (i.e., sets of compatible measurements) of a bipartite Bell experiment with binary choice of measurements on each side.

The settings also tell us whether or not we have "collapsed" the state of an agents' system and memory (or rest of the lab) by applying a projector corresponding to their outcome and provide a way to formalize Araujo's assumption. Araujo also noted issues with FR's treatment of postselection, which are accounted for in our analysis by explicitly computing probabilities conditioned on the postselection and avoiding collider bias. We have also seen that, in the FR scenario (Appendix G 2), the prediction $P(w = \text{fail} | r = \text{tails}, (x_1, x_2))$ does depend on the setting x_2 of F. This is analogous to the property that Healey calls intervention sensitivity.

While our results which establish the general consistency of quantum theory are certainly contrary to FR's popular summary that quantum theory cannot consistently justify the use of itself, the validity of FR's claimed theorem depends on how the assumptions are interpreted. In Sec. VC, we discussed a refined interpretation of FR's theorem in which it is correct; this would be to say that a version of quantum theory that additionally allows Heisenberg cuts to be freely ignored leads to contradictions with classical logic. However, in this interpretation, the result is not as surprising as the popular summary claims it to be, although it has undoubtedly fueled an intriguing research program on EWFSs in quantum foundations inspiring other no-go theorems (such as Refs. [6,7]) based on a similar setup where the underlying assumptions are formalized more rigorously.

APPENDIX J: PROOFS OF ALL RESULTS

1. Proofs of results from the main text

Lemma 1. If Σ is a set of consistent predictive statements, then

$$S \in \Sigma \Rightarrow \neg S \cap \Sigma = \emptyset, \quad (\text{J1})$$

where $\neg S$ denotes the negation of the statement S .

Proof. Let $S \in \Sigma$ be an arbitrary statement, by construction this of the form: "If the outcomes $\vec{a}_l$ take values $\vec{a}_l$ and the additional parameters of the scenario take the value $k = \ell$, then the outcomes $\vec{a}_l$ take values $\vec{a}_l$ with a probability $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, k = \ell)$." Then the negation of S is "If the outcomes $\vec{a}_l$ take values $\vec{a}_l$ and the additional parameters of the scenario take the value $k = \ell$, then the outcomes $\vec{a}_l$ do not take values $\vec{a}_l$ with a probability $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, k = \ell)$." More precisely, this is a set of statements.

$\neg S := \{\text{—If the outcomes } \vec{a}_l \text{ take values } \vec{a}_l \text{ and the additional parameters of the scenario take the value } k = \ell, \text{ then the outcomes } \vec{a}_l \text{ take values } \vec{a}_l \text{ with a probability } P'(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, k = \ell)\} \}_{P \neq P'}$.

It is then immediate from Definition 5 that if $S \in \Sigma$, then no $S' \in \neg S$ can be such that $S' \in \Sigma$. This completes the proof. ■

Corollary 1. If any two agents use the same choice of settings $\vec{x}$ for all measurements $\{\mathcal{M}^{A_i}\}_i$ in an augmented EWFS, then they make all the same predictions in that scenario.

Proof. The statement immediately follows from noting that fixing the setting choice for each measurement fully specifies the circuit in Fig. 2 and, therefore, the joint state of all systems and memories $S_1, \dots, S_m, M_1, \dots, M_N$ at each time step. If A and B make the same choice of settings for all measurements, then they fully agree on the circuit

(initial states and all channels) and, therefore, assign the same joint state to all systems at each time step.

Since all agents apply the Born rule to calculate the probabilities (Definition 8) and agree on the states and measurements for which the probabilities are calculated, they obtain the same probabilities and, therefore, make the same predictions for all measurements.

More explicitly, Appendix E shows that in any EWFS, given a choice of settings, every setting-conditioned prediction can be uniquely computed by applying the Born rule to our augmented circuit. Therefore, if all agents in an EWFS use the augmented quantum circuit to reason, picking the same setting choice or prior distribution $P(\vec{x})$, then they agree on all the predictions made in that scenario. ■

Corollary 2. If agents in an EWFS reason about each other's knowledge using the augmented circuit for the scenario, then they can never arrive at a logical contradiction even if they reason using all five assumptions **Q**, **U**, **C**, **D**, and **S**.

Proof. The proof proceeds by showing that all five assumptions are satisfied by statements Σ^{aug} derived in our augmented circuit framework, and the consistency of our framework shown in Appendix J 2 guarantees that all five can be consistently applied to reason without any contradictions.

Q and **U** are by construction satisfied for all the statements Σ^{aug} , since these are associated with predictions computed using the Born rule and consider all possible setting choices (including the unitary modeling). Moreover, all our results apply to general EWFS (Definition 1), which allows agents to have full quantum control over the labs of others, in the precise sense described in the formalization of **U**.

S holds for all statements Σ^{aug} in our framework, because the statements are obtained from setting-conditioned predictions, which are well-defined normalized probabilities (see Appendix E), and the set of statements Σ^{aug} is consistent as proven in Appendix J 2. No single valid normalized probability distribution can assign $P(\vec{a}_j = \vec{a}_j) = 1$ and $P(\vec{a}_j = \vec{a}'_j) = 1$, and the consistency result forbids the possibility of two distributions in the same scenario with $P(\vec{a}_j = \vec{a}_j) = 1$ and $P'(\vec{a}_j = \vec{a}'_j) = 1$.

C holds for all statements in our framework, because if an agent A_i knows that an agent A_j knows a statement $S \in \Sigma^{\text{aug}}$, then A_i can directly compute the setting-conditioned prediction $P(\vec{a}_j = \vec{a}_j | \vec{a}_i = \vec{a}_i, \vec{x} = \vec{\xi})$ that defines the statement S and thereby inherit the knowledge. Consistency as shown in Appendix J 2 guarantees that there is only one such probability assignment any user of our framework can arrive at for a given choice $\vec{x} = \vec{\xi}$ and outcome values $\vec{a}_j = \vec{a}_j$ and $\vec{a}_i = \vec{a}_i$ [85].

Finally, the distributive axiom **D** is a rather basic axiom of logical inference; that it holds in our framework can be

seen as follows. First, since $\Sigma_L^{\text{aug}} \subseteq \Sigma^{\text{aug}}$, consistency of the superset as shown in Appendix J 2 implies consistency of the subset. Then, consider a logical statement $S_1 \in \Sigma_L^{\text{aug}}$ associated with a logical setting-conditioned prediction $P(\vec{a}_j = \vec{a}_j | \vec{a}_i = \vec{a}_i, \vec{x} = \vec{\xi}) \in \{0, 1\}$; it can be of the following form where the set $\vec{a}_i$ of outcomes can be empty:

$$\begin{aligned} \vec{a}_i = \vec{a}_i \wedge \vec{x} = \vec{\xi} &\Rightarrow \neg(\vec{a}_j = \vec{a}_j), \\ \vec{a}_i = \vec{a}_i \wedge \vec{x} = \vec{\xi} &\Rightarrow \vec{a}_j = \vec{a}_j. \end{aligned} \quad (\text{J2})$$

Let S_2 be another statement of the same form, but relative to a potentially different set of outcomes $\vec{a}_m$ and $\vec{a}_n$ and setting values $\vec{x} = \vec{\xi}'$, i.e.,

$$\begin{aligned} \vec{a}_n = \vec{a}_n \wedge \vec{x} = \vec{\xi}' &\Rightarrow \neg(\vec{a}_m = \vec{a}_m), \\ \vec{a}_n = \vec{a}_n \wedge \vec{x} = \vec{\xi}' &\Rightarrow \vec{a}_m = \vec{a}_m. \end{aligned} \quad (\text{J3})$$

This corresponds to $P(\vec{a}_m = \vec{a}_m | \vec{a}_n = \vec{a}_n, \vec{x} = \vec{\xi}') \in \{0, 1\}$. Denote the probabilities associated with S_1 and S_2 as P_1 and P_2 in short. If $S_1 \Rightarrow S_2$, then P_2 can be derived from P_1 through the rules of classical probability theory and usual manipulation of quantum circuits (in this case for the augmented circuit).

This implies that, if one were to directly compute the probabilities for $\vec{a}_m$ and $\vec{a}_n$ under the setting choice $\vec{x} = \vec{\xi}'$ in the augmented circuit, one must arrive at P_2 (otherwise, there can be two distinct probability assignments to the same outcomes given these settings, which is not possible due to consistency, Definition 5, and Appendix J 2). This shows that $S_2 \in \Sigma_L^{\text{aug}}$. As this holds for all agents using our framework to reason, it follows that Eq. (C4) holds. ■

Theorem 1. There exists a consistent description of the FR protocol (both versions) that satisfies all five assumptions **Q**, **U**, **C**, **D**, and **S** but violates **I** for certain logical setting-conditioned predictions. Furthermore, when simultaneously assuming **Q**, **U**, **C**, **D**, and **S** in the FR protocol, additionally imposing **I** on at least one logical setting-conditioned prediction is a necessary condition for reproducing the apparent FR paradox, while imposing **I** on all logical setting-conditioned predictions is a sufficient condition for the same.

Proof. The existence claim follows from the main results of our framework Sec. IV and Corollary 2. The consistent description satisfying **Q**, **U**, **C**, **D**, and **S** is given by the augmented circuit. The violation of **I** for certain logical setting-conditioned prediction follows from the analysis of Appendix G 3, where we have shown that the logical predictions $P(b = 1 | u = w = \text{ok}, (x_1, x_2) = (0, 1)) = 1$ and $P(w = \text{fail} | a = 1, (x_1, x_2) = (1, 0)) = 1$ are setting dependent (thus violating **I**), since

$$P(b = 1|u = w = \text{ok}, (x_1, x_2) = (0, 1)) \neq P(b = 1|u = w = \text{ok}, (x_1, x_2) = (1, 1)),$$

$$P(w = \text{fail}|a = 1, (x_1, x_2) = (1, 0)) \neq P(w = \text{fail}|a = 1, (x_1, x_2) = (1, 1)). \quad (\text{J4})$$

In the prepare-and-measure version analyzed in Appendix G 2, the logical predictions $P(w = \text{fail}|r = \text{tails}, (x_1, x_2) = (1, 0)) = 1$ and $P(z = +\frac{1}{2}|\bar{w} = \bar{\text{ok}}, (x_1, x_2) = (0, 1)) = 1$ are setting dependent and violate **I**, since

$$P(w = \text{fail}|r = \text{tails}, (x_1, x_2) = (1, 0)) \neq P(w = \text{fail}|r = \text{tails}, (x_1, x_2) = (1, 1)),$$

$$P\left(z = +\frac{1}{2}|\bar{w} = \bar{\text{ok}}, (x_1, x_2) = (0, 1)\right) \neq P\left(z = +\frac{1}{2}|\bar{w} = \bar{\text{ok}}, (x_1, x_2) = (1, 1)\right). \quad (\text{J5})$$

These precisely correspond to the three statements $\bar{\text{F}}^{n:02}$ and $\bar{\text{W}}^{n:22}$ of FR that are combined to yield the apparent paradox (as shown in Table II).

The fact that assuming **I** is necessary to reproduce the apparent paradox in both versions of the protocol follows from Corollary 3, and the sufficiency follows from noting that assuming **I** for all logical setting-conditioned predictions allows us to ignore the setting information on all such predictions and, consequently, on all associated logical statements. It is clear from Tables I and II that this recovers the original statements of FR (in both versions) and, therefore, the apparent paradox. ■

Theorem 2 (nonaction on memory and setting independence). Consider an EWFS and a subset A_K of agents therein. Suppose that $A_i \notin A_K$ is another agent in the EWFS such that no agent in A_K acts as a superagent to A_i i.e., $(A_i, A_k) \in n\mathcal{SA} \ \forall \ A_k \in A_K$. Then, for every partition $A_K = \{A_{j_1}, \dots, A_{j_p}\} \cup \{A_{l_1}, \dots, A_{l_q}\}$ of A_K , the setting-conditioned prediction $P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi})$ is independent of the setting x_i ; that is,

$$P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi})$$

$$= P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, \vec{x} \setminus x_i = \vec{\xi} \setminus \xi_i), \quad (\text{J6})$$

where $\vec{x} \setminus x_i = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_N)$ is the setting vector with the i th component omitted, and similarly for the vector of setting values $\vec{\xi} \setminus \xi_i$. This expression is equivalent to the conditional independence given in Eq. (J10).

Proof. Given any set A_K of agents and another agent A_i such that $(A_i, A_k) \in n\mathcal{SA} \ \forall \ A_k \in A_K$, we first show that the joint probability of the outcomes of all agents in A_K (these outcomes are denoted using the vector $\vec{a}_K$) is independent of the setting x_i , i.e.,

$$P(\vec{a}_K|(x_1, \dots, x_N)) = P(\vec{a}_K|(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_N)). \quad (\text{J7})$$

Here, we have shortened the usual notion using value assignments $a = \mathcal{a}$ in probabilities to just the variable a , since the meaning of the equation is clear from context.

From this, the desired result follows easily through the rules of conditional probability.

To establish Eq. (J7), we consider the operationally equivalent augmented circuit (involving agents $\{A'_1, \dots, A'_N\}$) to the original augmented circuit (involving agents $\{A_1, \dots, A_N\}$), whose causal structure more explicitly reflects nonsuperagent structure $n\mathcal{SA}$ of the original EWFS. Recall that such an equivalent circuit is guaranteed to exist by Definition C8. The equivalence implies, in particular, the equivalence of predictions in the two circuits; that is, for all disjoint sets of outcomes $\vec{a}_j$ and $\vec{a}_l$ in the original EWFS,

$$P(\vec{a}_j = \vec{a}_j|\vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi}) = P(\vec{a}'_j = \vec{a}'_j|\vec{a}'_l = \vec{a}'_l, \vec{x}' = \vec{\xi}). \quad (\text{J8})$$

We establish Eq. (J7) for the primed EWFS and, by the above equivalence of predictions, obtain the same for the original EWFS. We do so by dividing the problem into different cases, depending on the causal structure of the (equivalent) augmented circuit. According to Definition C8, $(A_i, A_k) \in n\mathcal{SA}$ (along with the fact that we have $i \neq k$) implies that, in the equivalent EWFS, one of the following cases must hold:

- (i) *Case 1*— $A'_i \not\prec A'_k$;
- (ii) *Case 2*— $A'_i \prec A'_k$, and $\mathcal{BE}'_i$ acts trivially on M_i and $A_i \not\prec^{M_i} A_k$.

Case 1—In this case, Eq. (J7) immediately follows from the causality result of Appendix J 2.

Case 2—In this case, we must have the following property: For any agent A'_j with $A'_i \prec^{M_i} A'_j$, we must have $A'_j \not\prec A'_k$ for all K . Otherwise, $A'_i \prec^{M_i} A'_j$ along with the directed path $A'_j \prec A'_k$ would immediately imply by Definition C1 that $A'_i \prec^{M_i} A'_k$, which would contradict the condition of case 2 and, hence, the fact that $(A_i, A_k) \in n\mathcal{SA}$ (Definition C8).

Now, the prediction of interest from Eq. (J7) can be calculated explicitly by applying Eq. (E3) as detailed in Appendix E. The above property guarantees that, for all A'_j such that $A'_i \prec^{M_i} A'_j$, the operations of A'_j and those of any

agent $A'_k \in A'_K$ must act of disjoint sets of systems (Definition C1). Note that $A'_i \prec^{M_i} A'_j$ implies that A'_j acts after A'_i in the augmented circuit. Then, by the nonsignaling property of quantum theory, and as we can explicitly see from Eq. (E3), the trace in the probability calculation acts on all the outputs of A'_j 's operation. Since A'_j is not in the set A'_K , we sum over their outcomes in the probability; therefore, all operations of such agents are completely positive and trace preserving maps (CPTPMs). For CPTPMs, tracing the output after applying the map is equivalent to tracing the inputs directly $\text{tr} \circ \mathcal{B}E = \text{tr}$.

The above implies that all the operations of the agents A'_j such that $A'_i \xrightarrow{M_i} A'_j$ will drop out of the probability expression for Eq. (J7). This probability expression then has the trace over M_i acting directly on the operation $\mathcal{E}'_i \circ \mathcal{M}^{A_i}$ of the agent A'_i . From the definition of case 2, we also know that $\mathcal{E}'_i$ acts trivially on M_i , which implies that the trace over M_i acts directly on the measurement $\mathcal{M}^{A_i}$.

It is easy to see that $\text{tr}_{M_i} \circ \mathcal{M}^{A_i}$ is the same, independently of the setting $x_i \in \{0, 1\}$, because the setting dictates whether we coherently ($x_i = 0$) or incoherently ($x_i = 1$) copy the system state onto the memory in the basis of the measurement, and the postmeasurement state on the system alone (obtained by tracing out the memory) is the same in both cases. A more explicit proof of this fact can be found in the proof of Appendix J2. This is sufficient to establish the required Eq. (J7) for this case.

Having established Eq. (J7) for all sets A_K of agents of the form required by the theorem statement, it is immediate that the same setting independence of Eq. (J7) also holds for all subsets of agents A_K [this can be seen by computing the relevant marginals of Eq. (J7)]. Now, for any partition $A_K = \{A_{j_1}, \dots, A_{j_p}\} \cup \{A_{l_1}, \dots, A_{l_q}\}$ of A_K , we can compute prediction $P(\vec{a}_j | \vec{a}_l, \vec{x})$ by applying the conditional probability rule

$$P(\vec{a}_j | \vec{a}_l, \vec{x}) = \frac{P(\vec{a}_j, \vec{a}_l | \vec{x})}{P(\vec{a}_l | \vec{x})}. \quad (\text{J9})$$

The fact that both the numerator and denominator of this expression are independent of the component x_i of the setting vector $\vec{x}$ are then immediate from Eq. (J7). This establishes the theorem. ■

2. Proofs of results from the Appendixes

Theorem C1.

- (1) *Completeness*—In any given EWFS, all conventional predictions in that EWFS can be derived within the single augmented circuit of that EWFS. More explicitly, each conventional prediction $P_{\text{conv}}(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l)$ in the EWFS equals a particular setting conditioned prediction $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi}^*)$ of the augmented

circuit where the setting choice $\vec{x} = \vec{\xi}^*$ is such that $x_i = 1$ for all $i \in \{j_1, \dots, j_p, l_1, \dots, l_q\}$ and $x_i = 0$ for all $i \notin \{j_1, \dots, j_p, l_1, \dots, l_q\}$.

- (2) *Consistency*—For any EWFS, the set of all statements Σ^{aug} obtained in the corresponding augmented circuit (Definition 10) are consistent according to Definition 5.
- (3) *Causality*—For every setting-conditioned prediction $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{\xi})$ and every i such that $A_i \not\prec A_k$ for all $k \in \{j_1, \dots, j_p, l_1, \dots, l_q\}$, the prediction is independent of the setting x_i . That is, for all such i , we have the following, noting that $\vec{x} = \vec{\xi}$ is short for $x_1 = \xi_1, \dots, x_N = \xi_N$ and each $\xi_i \in \{0, 1\}$:

$$\begin{aligned} P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, (x_1 = \xi_1, \dots, x_i = 0, \dots, x_N = \xi_N)) \\ (= P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, (x_1 = \xi_1, \dots, x_i = 1, \dots, x_N = \xi_N))). \end{aligned} \quad (\text{J10})$$

Proof.

- (1) *Completeness*—Recall that a conventional prediction $P_{\text{conv}}(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l)$ in an EWFS (Definition 6) is computed by modeling the measurements $\mathcal{M}^{A_i}$ for all $i \in \{j_1, \dots, j_p, l_1, \dots, l_q\}$ as purely unitary evolutions of the agents' labs [Eq. (3)]. By construction of the augmented circuit for the EWFS, this corresponds to the case where $x_i = 0$ for all $i \in \{j_1, \dots, j_p, l_1, \dots, l_q\}$.

For $i \notin \{j_1, \dots, j_p, l_1, \dots, l_q\}$, in a conventional prediction, a projective measurement of $\{\Pi_{a_i}^{S_i} = |a_i\rangle\langle a_i|_{S_i}\}_{a_i \in \mathcal{O}_i}$ is performed followed by a CNOT in the same basis with the system S_i as control and memory M_i as target (capturing the memory update after measurement). This CNOT is precisely the unitary $\mathcal{M}_{\text{unitary}}^{A_i}$ [Eq. (3)].

Now, following Sec. III B in the main text, consider an initial state $|\psi\rangle_{S_i} \otimes |0\rangle_{M_i}$ of the system and memory (on which the measurement $\mathcal{M}^{A_i}$ acts) where $|\psi\rangle_{S_i} = \sum_{a_i \in \mathcal{O}_i} c_{a_i} |a_i\rangle_{S_i}$. That section of the main text shows that the following two procedures are operationally equivalent: (i) a projective measurement $\{\pi_{a_i}^{S_i} = |a_i\rangle\langle a_i|_{S_i}\}_{a_i \in \mathcal{O}_i}$ is applied on the system, and then the unitary channel $\mathcal{M}_{\text{unitary}}^{A_i}$ is applied on the system and memory; (ii) the unitary channel $\mathcal{M}_{\text{unitary}}^{A_i}$ is applied on the system and memory, and then the projective measurement $\{\pi_{a_i}^{S_i M_i} = |a_i a_i\rangle\langle a_i a_i|_{S_i M_i}\}_{a_i \in \mathcal{O}_i}$ is performed on the system and memory. The operational equivalence of (i) and (ii) entails that they yield the same transformation on the initial state and also that the

measurements in both cases yield the same probabilities for an outcome a_i .

Generally, by linearity, the argument about the equivalence of (i) and (ii) extends to all initial states $\rho_{S_i} \otimes |0\rangle\langle 0|_{M_i}$ of the system and memory. (i) is the procedure used for dealing with a measurement $\mathcal{M}^{A_i}$ when computing conventional predictions involving the outcome a_i (Definition 6), while (ii) is the procedure used in the augmented EWFS for calculating the setting conditioned predictions (Definition 8) involving the outcome a_i where by default we have $x_i = 1$.

This shows that the conventional prediction for the probability of $\vec{a}_j = \vec{a}_j$ given $\vec{a}_l = \vec{a}_l$ and the augmented circuit prediction for the same yield the same answer when using the setting assignment $\vec{x} = \vec{x}^*$ defined as $x_i = 1$ for $i \in \{j_1, \dots, j_p, l_1, \dots, l_q\}$ and $x_i = 0$ otherwise. The only difference is that the corresponding prediction in the augmented circuit explicitly conditions on the setting choice $\vec{x} = \vec{x}^*$, and we have $P_{\text{conv}}(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l) = P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{x}^*)$.

This establishes the claim that all the conventional predictions of any given EWFS can be derived as particular cases of setting-conditioned predictions in the single augmented circuit of the EWFS, showing the completeness of our formalism.

- (2) *Consistency*—The fact that the set Σ^{aug} of statements associated with an EWFS in our framework satisfies the consistency condition of Definition 5 immediately follows from the procedure through which these statements are derived.

As defined in Definition 10, each statement $S \in \Sigma^{\text{aug}}$ is associated with a setting-conditioned prediction $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{x})$ of the given EWFS. Such predictions are computed by applying the Born rule to a single, well-defined quantum circuit (explicitly detailed in Appendix E), which implies that, for any given setting choice $\vec{x} = \vec{x}$ and sets of outcome values $\vec{a}_j = \vec{a}_j$, $\vec{a}_l = \vec{a}_l$ in a given EWFS, a unique setting-conditioned prediction $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{x})$ can be computed, which corresponds to a valid, well-defined, and normalized conditional probability distribution.

Since Σ^{aug} contains only statements associated with setting-conditioned predictions in the augmented circuit and it is impossible to have $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{x})$ and $P'(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{x})$ in the same augmented EWFS for $P \neq P'$, the consistency of Σ^{aug} according to Definition 5 follows.

- (3) *Causality*—Our framework provides a single well-defined quantum circuit (the augmented circuit) from which all setting-conditioned predictions in an EWFS

can be derived and in which all the operations are applied in an acyclic order (given by the DAG G). From this, the desired result follows immediately by applying well-known results on quantum causal networks or causal models, such as the d -separation theorem [48,49,52,86]. However, in the interest of not introducing new concepts, we describe the proof in terms of concepts and results introduced in this paper.

Consider the setting-conditioned prediction $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{x})$ and the condition (C) $A_i \not\prec A_k$ for all $k \in \{j_1, \dots, j_p, l_1, \dots, l_q\}$. Notice that an A_i satisfies this condition, in particular, when $t_i > \max(t_{j_1}, \dots, t_{j_p}, t_{l_1}, \dots, t_{l_q})$. In Appendix E, we have established that any setting-conditioned prediction $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{x})$ in an augmented EWFS can be simplified to the form of Eq. (E6), where the components x_i of the setting vector $\vec{x}$ which are associated with a time $t_i > \max(j_1, \dots, j_p, l_1, \dots, l_q)$ do not feature in the probability expression, which implies the required independence Eq. (J10) for all such settings x_i .

For cases where we have $t_i < t_k$ for some $k \in \{j_1, \dots, j_p, l_1, \dots, l_q\}$ but $A_i \not\prec A_k$, by construction, this would happen only when the circuit contains no directed path of wires from the measurement $\mathcal{M}^{A_i}$ at t_i to the measurement $\mathcal{M}^{A_k}$ at $t_k > t_i$ (absence of directed paths between corresponding nodes in G). In this case, the two measurements act on disjoint sets of systems and the nonsignaling property of quantum theory would then guarantee the required independence of the outcome of one measurement from the setting of the other.

Alternatively, in such cases, we can always transform to an equivalent circuit (with the same channels, systems, states, and same connectivity between channels) where A_i and A_k are assigned time t'_i and t'_k with the opposite time order, $t'_i > t'_k$, allowing us to apply the argument from the previous paragraph to establish the required independence [87]. ■

Theorem F1 (recovering standard quantum circuits). If an EWFS corresponds to a standard quantum scenario (Definition 13), then its augmented circuit can be equivalently reduced to a standard quantum circuit, such that the same (nontrivial) predictions are obtained from the original augmented circuit, the C^{sys} -form standard circuit, or the $C^{\text{sys+anc}}$ -form standard circuit. Explicitly, for any disjoint sets $\vec{a}_j = (a_{j_1}, \dots, a_{j_p})$ and $\vec{a}_l = (a_{l_1}, \dots, a_{l_q})$ of outcomes, and any choice of settings $\vec{x} = \vec{x}$ such that $x_i = 1$ for all $i \in \{j_1, \dots, j_p, l_1, \dots, l_q\}$, we have

$$P_{\text{aug}}(\vec{a}_j = \vec{a}_j | \vec{a}_j = \vec{a}_j, \vec{x} = \vec{x}) = P_{\text{std}}(\vec{a}'_j = \vec{a}_j | \vec{a}'_j = \vec{a}_j), \quad (\text{J11})$$

where the P_{aug} refers to setting-conditioned predictions in the augmented circuit of the EWFS and P_{std} refers to predictions in an equivalent $\mathcal{C}^{\text{sys}}$ -form or $\mathcal{C}^{\text{sys}+\text{anc}}$ -form standard quantum circuit (where no settings are involved).

Proof. In a standard quantum scenario, by Definition 13, we must have $(A_i, A_j) \in n\mathcal{SA}$ for all $i, j \in \{1, \dots, N\}$. This means that the augmented circuit of the given EWFS over agents $\{A_1, \dots, A_N\}$ can be reduced to an operationally equivalent augmented circuit over corresponding agents $\{A'_1, \dots, A'_N\}$ such that for any A'_i and A'_j we have either A'_j acts before A'_i in time $t'_j > t'_i$ or, in the alternative cases where $i = j$ or $t'_j > t'_i$, no operation that acts after the measurement $\mathcal{M}^{A'_i}$ in the circuit (including the operation $\mathcal{E}'_i$) acts nontrivially on the memory M'_i . Here, we have used the fact that $A'_i < A'_j$ according to the operational causal structure (Definition C1) implies $t'_i < t'_j$.

Applying this argument to every pair of agents in the scenario, this implies that, for each measurement $\mathcal{M}^{A'_i}$, the corresponding set of systems S'_i of systems (excluding the memory M'_i) on which it acts nontrivially is a subset of the systems S' and does not include any of the memories $M'_j \in M'$ of other agents (in contrast to a general EWFS of Definition 1, where the “system” for one agent’s

measurement may include the “memories” of other agents). All the following arguments refer to the equivalent augmented circuit over the primed agents.

Let the measurement outcomes a_i of A'_i ’s measurement take values in the same set $\mathcal{O}_i$ as that of the original scenario without loss of generality [88]. Then, A'_i ’s measurement on the systems S'_i is associated with the projectors $\{|a_i\rangle\langle a_i|_{S'_i}\}_{a_i \in \mathcal{O}_i}$. Let $\rho_{S'_i}$ be an arbitrary state, which corresponds to the state of the system S'_i just before A'_i ’s measurement. We can express this state in the basis of the measurement as follows, for some coefficients $c_{a_i, \bar{a}_i}$:

$$\rho_{S'_i} = \sum_{a_i, \bar{a}_i} c_{a_i, \bar{a}_i} |a_i\rangle\langle \bar{a}_i|_{S'_i},$$

where $\{|a_i\rangle_i\}$ and $\{|\bar{a}_i\rangle_i\}$ both correspond to the same measurement basis. Recall that the memory M'_i is initialized to $|0\rangle\langle 0|_{M'_i}$. Now, consider the case where the setting $x'_i = 0$. Then, we model A'_i ’s measurement as the unitary evolution $\mathcal{M}^{A'_i}_{\text{unitary}}$ which, as we have seen before, is simply a unitary implementing a coherent copy in the measurement basis with S'_i being the control and M'_i being the target. The postmeasurement state of $S'_i M'_i$ in this case is given as

$$\mathcal{M}^{A'_i}_{\text{unitary}}(\rho_{S'_i} \otimes |0\rangle\langle 0|_{M'_i}) \mathcal{M}^{A'_i, \dagger}_{\text{unitary}} = \sum_{a_i, \bar{a}_i} c_{a_i, \bar{a}_i} |a_i a_i\rangle\langle \bar{a}_i \bar{a}_i|_{S'_i M'_i}. \quad (\text{J12})$$

Then it is easy to see that the following holds, where we use the notation $\pi^{S'_i}_{a_i} := |a_i\rangle\langle a_i|_{S'_i}$, $\pi^{M'_i}_{\bar{a}_i} := |\bar{a}_i\rangle\langle \bar{a}_i|_{M'_i}$, $\pi^{S'_i M'_i}_{a_i \bar{a}_i} = |a_i a_i\rangle\langle \bar{a}_i \bar{a}_i|_{S'_i M'_i}$:

$$(1_{S'_i} \otimes \pi^{M'_i}_{\bar{a}_i})(\mathcal{M}^{A'_i}_{\text{unitary}}(\rho_{S'_i} \otimes |0\rangle\langle 0|_{M'_i}) \mathcal{M}^{A'_i, \dagger}_{\text{unitary}})(1_{S'_i} \otimes \pi^{M'_i}_{\bar{a}_i}) = \pi^{S'_i M'_i}_{a_i \bar{a}_i} \mathcal{M}^{A'_i}_{\text{unitary}}(\rho_{S'_i} \otimes |0\rangle\langle 0|_{M'_i}) \mathcal{M}^{A'_i, \dagger}_{\text{unitary}} \pi^{S'_i M'_i}_{a_i \bar{a}_i}. \quad (\text{J13})$$

We now show that the postmeasurement on the system S'_i alone is independent of the setting x'_i , and this is indeed the postmeasurement state one would get by directly applying the projective measurement of $\mathcal{C}^{\text{sys}}$ on S'_i . This allows us to reduce our augmented circuit to a $\mathcal{C}^{\text{sys}}$ -form standard quantum circuit as required. For this, first consider this state under the setting $x'_i = 0$. This is given as $\text{tr}_{M'_i}[\mathcal{M}^{A'_i}_{\text{unitary}}(\rho_{S'_i} \otimes |0\rangle\langle 0|_{M'_i}) \mathcal{M}^{A'_i, \dagger}_{\text{unitary}}]$. Noting that $\sum_{a_i \in \mathcal{O}_i} \pi^{M'_i}_{\bar{a}_i} = 1_{M'_i}$, we can expand this as follows:

$$\begin{aligned} & \text{tr}_{M'_i}[\mathcal{M}^{A'_i}_{\text{unitary}}(\rho_{S'_i} \otimes |0\rangle\langle 0|_{M'_i}) \mathcal{M}^{A'_i, \dagger}_{\text{unitary}}] \\ &= \sum_{\bar{a}_i \in \mathcal{O}_i} (1_{S'_i} \otimes \langle \bar{a}_i|_{M'_i})(\mathcal{M}^{A'_i}_{\text{unitary}}(\rho_{S'_i} \otimes |0\rangle\langle 0|_{M'_i}) \mathcal{M}^{A'_i, \dagger}_{\text{unitary}})(1_{S'_i} \otimes |\bar{a}_i\rangle_{M'_i}) \\ &= \sum_{\bar{a}_i \in \mathcal{O}_i} (1_{S'_i} \otimes \langle \bar{a}_i|_{M'_i}) \left(1_{S'_i} \otimes \sum_{a_i \in \mathcal{O}_i} \pi^{M'_i}_{\bar{a}_i} \right) (\mathcal{M}^{A'_i}_{\text{unitary}}(\rho_{S'_i} \otimes |0\rangle\langle 0|_{M'_i}) \mathcal{M}^{A'_i, \dagger}_{\text{unitary}}) \left(1_{S'_i} \otimes \sum_{a_i \in \mathcal{O}_i} \pi^{M'_i}_{\bar{a}_i} \right) (1_{S'_i} \otimes |\bar{a}_i\rangle_{M'_i}) \\ &= \sum_{\bar{a}_i \in \mathcal{O}_i} \sum_{a_i \in \mathcal{O}_i} (1_{S'_i} \otimes \langle \bar{a}_i|_{M'_i})(1_{S'_i} \otimes \pi^{M'_i}_{\bar{a}_i})(\mathcal{M}^{A'_i}_{\text{unitary}}(\rho_{S'_i} \otimes |0\rangle\langle 0|_{M'_i}) \mathcal{M}^{A'_i, \dagger}_{\text{unitary}})(1_{S'_i} \otimes \pi^{M'_i}_{\bar{a}_i})(1_{S'_i} \otimes |\bar{a}_i\rangle_{M'_i}) \\ &= \text{tr}_{M'_i} \left[\sum_{a_i \in \mathcal{O}_i} (1_{S'_i} \otimes \pi^{M'_i}_{\bar{a}_i})(\mathcal{M}^{A'_i}_{\text{unitary}}(\rho_{S'_i} \otimes |0\rangle\langle 0|_{M'_i}) \mathcal{M}^{A'_i, \dagger}_{\text{unitary}})(1_{S'_i} \otimes \pi^{M'_i}_{\bar{a}_i}) \right]. \end{aligned} \quad (\text{J14})$$

In the second equality above, we have summed over the same indices $\sum_{a_i \in \mathcal{O}_i}$ in both occurrences of the projectors $\pi_{a_i}^{M'_i}$ as the cross terms disappear due to the trace over M_i . Using this and Eq. (J13), we immediately obtain the following:

$$\begin{aligned} & \text{tr}_{M'_i} \left[\mathcal{M}_{\text{unitary}}^{A'_i} (\rho_{S'_i} \otimes |0\rangle\langle 0|_{M'_i}) \mathcal{M}_{\text{unitary}}^{A'_i, \dagger} \right] \\ &= \text{tr}_{M'_i} \left[\sum_{a_i \in \mathcal{O}_i} \pi_{a_i}^{S'_i M'_i} \mathcal{M}_{\text{unitary}}^{A'_i} (\rho_{S'_i} \otimes |0\rangle\langle 0|_{M'_i}) \mathcal{M}_{\text{unitary}}^{A'_i, \dagger} \pi_{a_i}^{S'_i M'_i} \right] \\ &= \sum_{a_i \in \mathcal{O}_i} c_{a_i, a_i} |a_i\rangle\langle a_i|_{S'_i}. \end{aligned} \quad (\text{J15})$$

The right-hand side is precisely the expression for the postmeasurement state of S'_i under the setting $x_i = 1$ when using the trace-preserving form of the evolution for this setting [cf. Eq. (9)]. Thus, we have shown that A'_i 's measurement implements the following map on an arbitrary input state $\rho_{S'_i}$ of the system S'_i irrespective of the setting x'_i :

$$\rho_{S'_i} = \sum_{a_i, \bar{a}_i \in \mathcal{O}_i} c_{a_i, \bar{a}_i} |a_i\rangle\langle \bar{a}_i|_{S'_i} \mapsto \sum_{a_i \in \mathcal{O}_i} c_{a_i, a_i} |a_i\rangle\langle a_i|_{S'_i}. \quad (\text{J16})$$

This map can equivalently be described as

$$\rho_{S'_i} \mapsto \sum_{a_i \in \mathcal{O}_i} \pi_{a_i}^{S'_i} (\rho_{S'_i}) \pi_{a_i}^{S'_i}. \quad (\text{J17})$$

Or, for the case of a particular outcome, we have the following trace nonincreasing map from pre- to postmeasurement state:

$$\rho_{S'_i} \mapsto \frac{\pi_{a_i}^{S'_i} (\rho_{S'_i}) \pi_{a_i}^{S'_i}}{\text{tr}[\pi_{a_i}^{S'_i} (\rho_{S'_i}) \pi_{a_i}^{S'_i}]}. \quad (\text{J18})$$

These correspond precisely to how the measurements $\mathcal{M}^{A'_i} = \{\pi_{a_i}^{S'_i}\}_{a_i \in \mathcal{O}'_i}$ of a $\mathcal{C}^{\text{sys}}$ -form circuit (Definition F1) over $\{A'_1, \dots, A'_N\}$ act.

To show the full reduction of the augmented circuit to a $\mathcal{C}^{\text{sys}}$ -form circuit, let $S'_i{}^c := S' \setminus S'_i$ denote the complement of $S'_i \subseteq S'$ (the systems that $\mathcal{M}^{A'_i}$ acts nontrivially on). The above argument immediately generalizes to the case where we consider $\rho_{S'_i}$ to be a reduced state of a larger state $\rho_{S'_i{}^c S'_i}$ over all systems S' in the EWFS. Since under both settings the measurement operation of A'_i acts only locally on S'_i by construction, the above argument implies that the joint state on $S'_i{}^c S'_i$ is also independent of the settings. It follows that, for all agents A'_i , we can replace their measurement in the (primed) augmented circuit with the setting-independent projective measurement $\{|a_i\rangle\langle a_i|_{S'_i}\}_{a_i \in \mathcal{O}'_i}$ on the system $S'_i \subseteq S'$ alone while preserving all the predictions; i.e., the

augmented circuit reduces to an equivalent $\mathcal{C}^{\text{sys}}$ -form standard circuit. Since the primed circuit is operationally equivalent to the original augmented EWFS, it follows that the $\mathcal{C}^{\text{sys}}$ -form standard circuit obtained here is also operationally equivalent to the original augmented circuit.

To obtain an equivalent $\mathcal{C}^{\text{sys+anc}}$ -form circuit, we can keep the memories (even if they are not acted upon after A'_i 's measurement) and model all measurements with $x'_i = 1$ (recalling that we have established full setting independence of predictions, therefore an arbitrary fixing of settings will not change the predictions). Then, it follows from Eq. (J13) that this is equivalent to modeling each measurement in its unitary form $\mathcal{M}_{\text{unitary}}^{A'_i}$ at the time t'_i of the measurement followed by a projective measurement $\{\pi_{a_i}^{M'_i}\}_{a_i \in \mathcal{O}_i}$ on the memory M_i alone. Since the memory is not subsequently acted upon until the time t'_N at which the protocol ends, the measurement $\{\pi_{a_i}^{M'_i}\}_{a_i \in \mathcal{O}_i}$ acting on the memory alone can be equivalently performed at any time t_f after t'_N . This immediately yields $\mathcal{C}^{\text{sys+anc}}$ -form standard circuit which is equivalent to the original augmented circuit, where the memories $\{M'_1, \dots, M'_N\}$ act as the ancillas. This completes the proof of the current theorem. ■

-
- [1] E. P. Wigner, *Remarks on the Mind-Body Question* (Indiana University Press, Bloomington, 1967), pp. 171–184, https://link.springer.com/chapter/10.1007/978-3-642-78374-6_20.
 - [2] Veronika Baumann and Stefan Wolf, *On formalisms and interpretations*, *Quantum* **2**, 99 (2018).
 - [3] Daniela Frauchiger and Renato Renner, *Quantum theory cannot consistently describe the use of itself*, *Nat. Commun.* **9**, 3711 (2018).
 - [4] Nuriya Nurgalieva and Lidia del Rio, *Inadequacy of modal logic in quantum settings*, *EPTCS* **287**, 267 (2019).
 - [5] Časlav Brukner, *On the quantum measurement problem*, *arXiv:1507.05255*.
 - [6] Časlav Brukner, *A no-go theorem for observer-independent facts*, *Entropy* **20**, 350 (2018).
 - [7] Kok-Wei Bong, Aníbal Utreras-Alarcón, Farzad Ghafari, Yeong-Cherng Liang, Nora Tischler, Eric G. Cavalcanti, Geoff J. Pryde, and Howard M. Wiseman, *A strong no-go theorem on the Wigner's friend paradox*, *Nat. Phys.* **16**, 1199 (2020).
 - [8] Howard M. Wiseman, Eric G. Cavalcanti, and Eleanor G. Rieffel, *A “thoughtful” local friendliness no-go theorem: A prospective experiment with new assumptions to suit*, *Quantum* **7**, 1112 (2023).
 - [9] William J. Zeng, Farrokh Labib, and Vincent Russo, *Towards violations of local friendliness with quantum computers*, *Quantum* **9**, 1851 (2025).
 - [10] Matthew Pusey, *An inconsistent friend*, *Nat. Commun.* **14**, 977 (2018).
 - [11] Nuriya Nurgalieva and Renato Renner, *Testing quantum theory with thought experiments*, *Contemp. Phys.* **61**, 193 (2020).

- [12] Joseph M. Renes, *Consistency in the description of quantum measurement: Quantum theory can consistently describe the use of itself*, [arXiv:2107.02193](https://arxiv.org/abs/2107.02193).
- [13] Marcelo Losada, Roberto Laura, and Olimpia Lombardi, *Frauchiger-Renner argument and quantum histories*, *Phys. Rev. A* **100**, 052114 (2019).
- [14] Alexios P. Polychronakos, *Quantum mechanical rules for observed observers and the consistency of quantum theory*, *Nat. Commun.* **15**, 3023 (2024).
- [15] Scott Aaronson, *It's hard to think when someone Hada-mards your brain*, <https://scottaaronson.blog/?p=3975> (2018).
- [16] Richard Healey, *Quantum theory and the limits of objec-tivity*, *Found. Phys.* **48**, 1568 (2018).
- [17] Mateus Araújo, *The flaw in Frauchiger, and Renner's argument*, <https://mateusaraujo.info/2018/10/24/the-flaw-in-frauchiger-and-renners-argument/> (2018).
- [18] Anthony Sudbery, *The hidden assumptions of Frauchiger and Renner*, *Int. J. Quantum Found.* **5**, 98 (2019), <https://ijqf.org/archives/5470>.
- [19] Eric G. Cavalcanti and Howard M. Wiseman, *Implications of local friendliness violation for quantum causality*, *Entropy* **23**, 925 (2021).
- [20] Yilè Ying, Marina Maciel Ansanelli, Andrea Di Biagio, Elie Wolfe, David Schmid, and Eric Gama Cavalcanti, *Relating Wigners friend scenarios to nonclassical causal compati-bility, monogamy relations, and fine tuning*, *Quantum* **8**, 1485 (2024).
- [21] Lúcia del Rio and Renato Renner, *Reply to: Quantum mechanical rules for observed observers and the consis-tency of quantum theory*, *Nat. Commun.* **15**, 3024 (2024).
- [22] V. Vilasini and Mischa P. Woods (to be published).
- [23] David Deutsch, *Quantum theory as a universal physical theory*, *Int. J. Theor. Phys.* **24**, 1 (1985).
- [24] In the following equation, we have used a pure initial state $|\psi\rangle_{S_i}$ to make the equations more concise, but it is easy to verify that analogous equations and the same arguments hold for mixed initial states ρ_{S_i} .
- [25] Collapse theories, for instance, would violate this assumption and prescribe different probabilities and rules for computing them.
- [26] One can formally write down the probability expressions obtained in steps 3 and 4 through the Born rule. However, in the interests of conciseness, we decline from doing so at this point, as we show that these conventional predictions since are recovered as a special case of our to-be-described and rigorously defined general formalism (Appendix J 2).
- [27] FR refer to C as an interagent “consistency” assumption, hence the name. Following the formulation of this in terms of epistemic modal logic [4,28], we find it more appropriate to view it as an inheritance of agents’ knowledge. However, we stick to the original acronym for simplicity. D refers to the distributive axiom of logic, and S corresponds to an assumption about measurements yielding “single outcomes” in every experimental run.
- [28] V. Vilasini, Nuriya Nurgalieva, and Lúcia del Rio, *Multi-agent paradoxes beyond quantum theory*, *New J. Phys.* **21**, 113028 (2019).
- [29] Recall that, in our framework, the settings $x_i = 1$ for all outcomes a_i belong to the outcome sets $\vec{a}_j$ or $\vec{a}_l$ appearing in the given prediction $P(\vec{a}_j = \vec{a}_j | \vec{a}_l = \vec{a}_l, \vec{x} = \vec{x})$; the re-maining components of $\vec{x}$ can be varied.
- [30] This aligns with a previous work, which also argued that FR’s conclusions hold only under a specific way of applying quantum theory [31]. However, our formalisation of this point via the assumption **I** allows us to further achieve a full resolution of FR-type paradoxes meeting the desiderata outlined in Sec. I A.
- [31] Anthony Sudbery, *Single-world theory of the extended Wigner’s friend experiment*, *Found. Phys.* **47**, 658 (2017).
- [32] We need not consider the case where Ursula and Wigner’s measurements are modeled unitarily (setting 0), as this is the case only if there were further super-superagents who measured Ursula and Wigner’s labs.
- [33] See Ref. [34] for further details on a general, theory-independent formalisation of this class of Wigner’s friend (apparent) paradoxes, derivable only by combining state-ments about compatible measurements, in addition to imposing **I**, which we have shown to be necessary.
- [34] Nuriya Nurgalieva and V. Vilasini, *Any theory that admits an extended Wigner’s friend-type paradox is logically contextual*, [arXiv:2502.03874](https://arxiv.org/abs/2502.03874).
- [35] Keep in mind that each agent in our formalism is associated with a single time step, and a physical agent acting at many time steps would be modeled as multiple agents acting here; thus, one agent has to act later in time than another in order to act on the latter’s memory.
- [36] Caroline L. Jones and Markus P. Müller, *On the significance of Wigner’s friend in contexts beyond quantum foundations*, [arXiv:2402.08727](https://arxiv.org/abs/2402.08727).
- [37] These structures and the setting independences they entail follow only from the properties of the EWFS protocol (and not agents’ subjective choices of cuts) and can be included in the initial protocol specification given to the agents.
- [38] Robert B. Griffiths, *Consistent histories and the inter-pretation of quantum mechanics*, *J. Stat. Phys.* **36**, 219 (1984).
- [39] Robert B. Griffiths, *The consistent histories approach to quantum mechanics*, in *The Stanford Encyclopedia of Philosophy (Summer 2024 Edition)*, edited by Edward N. Zalta and Uri Nodelman (Stanford, Stanford, CA, 2024), <https://plato.stanford.edu/entries/qm-consistent-histories>.
- [40] Marek Żukowski and Marcin Markiewicz, *Physics and metaphysics of Wigner’s friends: Even performed premea-surements have no results*, *Phys. Rev. Lett.* **126**, 130402 (2021).
- [41] This holds even when no projection postulate is used to set postmeasurement states; see Sec. V B.
- [42] C. G. Knott, *Life and Scientific Work of Peter Guthrie Tait* (Cambridge University Press, Cambridge, England, 1911); Reproduced in *The Scientific Letters and Papers of James Clerk Maxwell II: 1862–1873*, edited by P. M. Harman (Cambridge University Press, Cambridge, England, 1995).
- [43] Rolf Landauer, *Irreversibility and heat generation in the computing process*, *IBM J. Res. Dev.* **5**, 183 (1961).
- [44] Charles H. Bennett, *The thermodynamics of computation—A review*, *Int. J. Theor. Phys.* **21**, 905 (1982).
- [45] Commonly called indefinite causal order [46,47], though this terminology is often debated.

- [46] Ognyan Oreshkov, Fabio Costa, and Časlav Brukner, *Quantum correlations with no causal order*, *Nat. Commun.* **3**, 1092 (2012).
- [47] G. Chiribella, G. M. D’Ariano, P. Perinotti, and B. Valiron, *Quantum computations without definite causal structure*, *Phys. Rev. A* **88**, 022318 (2013).
- [48] Judea Pearl, *Causality: Models, Reasoning, and Inference* 2nd ed. (Cambridge University Press, Cambridge, England, 2009), <https://www.cambridge.org/core/books/causality/B0046844FAE10CBF274D4ACBDAEB5F5B>.
- [49] Joe Henson, Raymond Lal, and Matthew F. Pusey, *Theory-independent limits on correlations from generalized Bayesian networks*, *New J. Phys.* **16**, 113043 (2014).
- [50] Fabio Costa and Sally Shrapnel, *Quantum causal modelling*, *New J. Phys.* **18**, 063032 (2016).
- [51] John-Mark A. Allen, Jonathan Barrett, Dominic C. Horsman, Ciarán M. Lee, and Robert W. Spekkens, *Quantum common causes and quantum causal models*, *Phys. Rev. X* **7**, 031021 (2017).
- [52] Jonathan Barrett, Robin Lorenz, and Ognyan Oreshkov, *Quantum causal models*, [arXiv:1906.10726](https://arxiv.org/abs/1906.10726).
- [53] Lucien Hardy, *Probability theories with dynamic causal structure: A new framework for quantum gravity*, [arXiv:gr-qc/0509120](https://arxiv.org/abs/gr-qc/0509120).
- [54] Mateus Araújo, Philippe Allard Guérin, and Ämin Baumeler, *Quantum computation with indefinite causal structures*, *Phys. Rev. A* **96**, 052315 (2017).
- [55] Jonathan Barrett, Robin Lorenz, and Ognyan Oreshkov, *Cyclic quantum causal models*, *Nat. Commun.* **12**, 885 (2021).
- [56] Carla Ferradini, Victor Gitton, and V. Vilasini, *Cyclic functional causal models beyond unique solvability with a graph separation theorem*, [arXiv:2502.04171](https://arxiv.org/abs/2502.04171).
- [57] Carla Ferradini, Victor Gitton, and V. Vilasini, *Cyclic quantum causal modelling with a graph separation theorem*, [arXiv:2502.04168](https://arxiv.org/abs/2502.04168).
- [58] V. Vilasini and R. Colbeck, *General framework for cyclic and fine-tuned causal models and their compatibility with space-time*, *Phys. Rev. A* **106**, 032204 (2022).
- [59] V. Vilasini and R. Colbeck, *Impossibility of superluminal signaling in Minkowski spacetime does not rule out causal loops*, *Phys. Rev. Lett.* **129**, 110401 (2022).
- [60] V. Vilasini and R. Renner, *Embedding cyclic information-theoretic structures in acyclic space-times: No-go results for indefinite causality*, *Phys. Rev. A* **110**, 022227 (2024).
- [61] V. Vilasini and R. Renner, *Fundamental limits for realizing quantum processes in spacetime*, *Phys. Rev. Lett.* **133**, 080201 (2024).
- [62] Nick Ormrod, V. Vilasini, and Jonathan Barrett, *Which theories have a measurement problem?*, *New J. Phys.* **28**, 024506 (2026).
- [63] Sandu Popescu and Daniel Rohrlich, *Quantum nonlocality as an axiom*, *Found. Phys.* **24**, 379 (1994).
- [64] The memory M_i of an agent A_i is chosen to be a system of at least the same dimension as the system S_i that the agent A_i measures.
- [65] Asher Peres, *Quantum Theory: Concepts and Methods*, Fundamental Theories of Physics Vol. 72 (Kluwer Academic Publishers, New York, 1993), 10.1007/0-306-47120-5.
- [66] In simple terms, a DAG is a graph with additional structure: (i) The edges have a direction associated with them. (ii) By “following the edges in the indicated direction,” one can never get back to the starting point, i.e., no “directed loops.”
- [67] In Ref. [4], the assumption C is replaced by what they call the trust axiom, which asserts that an agent A^i can inherit the knowledge of another agent A^j as per Eq. (C3) only if A^i trusts A^j . In the FR setup, it is precisely these pairs of agents who can be said to trust each other. Other pairs such as Alice and Ursula need not trust each other, as one agent Hadamard’s the memory of the other.
- [68] Christopher J. Wood and Robert W. Spekkens, *The lesson of causal discovery algorithms for quantum correlations: Causal explanations of Bell-inequality violations require fine-tuning*, *New J. Phys.* **17**, 033002 (2015).
- [69] Armando Relaño, *Decoherence allows quantum theory to describe the use of itself*, [arXiv:1810.07065](https://arxiv.org/abs/1810.07065).
- [70] Armando Relaño, *Decoherence framework for Wigner’s-friend experiments*, *Phys. Rev. A* **101**, 032107 (2020).
- [71] Ruth E. Kastner, *Unitary-only quantum theory cannot consistently describe the use of itself: On the Frauchiger–Renner paradox*, *Found. Phys.* **50**, 441 (2020).
- [72] Andrea Di Biagio and Carlo Rovelli, *Stable facts, relative facts*, *Found. Phys.* **51**, 30 (2021).
- [73] Jeffrey Bub, *‘Two dogmas’ redux*, [arXiv:1907.06240](https://arxiv.org/abs/1907.06240).
- [74] Varun Narasimhachar, *Agents governed by quantum mechanics can use it intersubjectively and consistently*, [arXiv:2010.01167](https://arxiv.org/abs/2010.01167).
- [75] Aurélien Drezet, *About Wigner Friend’s and Hardy’s paradox in a Bohmian approach: A comment of ‘Quantum theory cannot’ consistently describe the use of itself*, [arXiv:1810.10917](https://arxiv.org/abs/1810.10917).
- [76] Sebastian Fortin and Olimpia Lombardi, *Wigner and his many friends: A new no-go result?*, [arXiv:1904.07412](https://arxiv.org/abs/1904.07412).
- [77] The cited works are only representatives of the extensive research generated by the FR paper. For further references on previous responses to FR, see Ref. [11].
- [78] David Schmid, Yilè Yīng, and Matthew Leifer, *A review and analysis of six extended Wigner’s friend arguments*, [arXiv:2308.16220](https://arxiv.org/abs/2308.16220).
- [79] Carlo Rovelli, *Relational quantum mechanics*, *Int. J. Theor. Phys.* **35**, 1637 (1996).
- [80] C. M. Caves, C. A. Fuchs, and R. Schack, *Quantum probabilities as Bayesian probabilities*, *Phys. Rev. A* **65**, 022305 (2002).
- [81] Christopher A. Fuchs, N. David Mermin, and Rüdiger Schack, *An introduction to QBism with an application to the locality of quantum mechanics*, *Am. J. Phys.* **82**, 749 (2014).
- [82] Stuart Samuel, *The Frauchiger-Renner gedanken experiment: An interesting laboratory for exploring some topics in quantum mechanics*, [arXiv:2208.00060](https://arxiv.org/abs/2208.00060).
- [83] Hugh Everett, *“Relative state” formulation of quantum mechanics*, *Rev. Mod. Phys.* **29**, 454 (1957).
- [84] Although additional steps can be employed to simplify the reasoning by dropping redundant settings and parameters, as is also the case in classical multiagent scenarios (cf. Sec. VII A).
- [85] This fact is independent of whether or not different agents agree on the setting choices to model their

perspective of the experiment, as, for instance, even if Wigner models the friend's lab as a unitarily evolving closed quantum system ($x_F = 0$), both Wigner and the friend can still use our augmented circuit to compute predictions for the case where $x_F = 1$ and will arrive at the same answer.

- [86] Robert R. Tucci, *Factorization of quantum density matrices according to Bayesian and Markov networks*, [arXiv:quant-ph/0701201](#).
- [87] Physically, if we regard the circuit as embedded in space-time such that the absence of directed paths between A_i and A_k correspond to spacelike separation, then the above transformation can be seen as transforming to another reference frame where the time order relative to the coordinate time is reversed.
- [88] Since the scenarios are operationally equivalent with a one-one correspondence between the outcome sets, we can use the same labels without loss of generality.